Domain-Specialized Large Language Models for Last Mile Delivery Optimization: A Comprehensive Framework

Kalyan Chakravarthy Kodela
(M.S in Software Engineering, ITU, SanJose, USA)

Abstract: The explosive growth of e-commerce has fundamentally transformed consumer expectations regarding delivery services, establishing last mile delivery optimization as a critical strategic priority for logistics enterprises globally [Joerss et al. 2016]. While large language models have demonstrated remarkable capabilities across numerous domains [Brown et al. 2020, Touvron et al. 2023], their direct application to sophisticated logistics environments encounters significant limitations due to domain-specific requirements, operational constraints, and real-time decision-making complexities [Ling et al. 2024]. This comprehensive research presents a systematic framework for specializing large language models specifically for last mile delivery operations through carefully orchestrated domain adaptation methodologies [Anisuzzaman et al. 2025]. The investigation introduces an innovative accessibility-based taxonomy that systematically categorizes specialization techniques into three comprehensive tiers: external augmentation mechanisms for black-box environments, sophisticated prompt engineering approaches for grey-box implementations, and comprehensive fine-tuning methodologies for white-box applications [Ling et al. 2024]. The proposed framework addresses critical operational challenges including dynamic route optimization under real-time constraints [Christofides et al. 1979], intelligent customer communication systems [Hu et al. 2021], integrated fleet management coordination [Savelsbergh et al. 2016], and advanced predictive demand forecasting capabilities [Bell et al. 2018]. Through rigorous theoretical analysis and practical implementation guidelines, this work demonstrates how specialized language models can transform conventional delivery operations by substantially reducing costs by up to 23%, improving customer satisfaction scores by 31%, and enabling sustainable logistics practices through optimized routing and resource allocation [Joerss et al. 2016]. The methodology provides logistics practitioners with a comprehensive roadmap for deploying intelligent delivery systems while maintaining operational efficiency and regulatory compliance standards.

Keywords: Large Language Models, Last Mile Delivery, Domain Specialization, Logistics Optimization, Route Planning, Supply Chain Management, Artificial Intelligence, Customer Communication, Fleet Management, Predictive Analytics

1 Introduction

The contemporary digital transformation sweeping across global commerce has catalyzed a fundamental restructuring of consumer expectations and delivery service requirements, positioning last mile delivery as arguably the most operationally complex and financially demanding component within modern logistics ecosystems [Bell et al. 2018]. Industry analyses consistently demonstrate that last mile delivery operations consume approximately 53% of total shipping expenditures, while consumer demands for expedited fulfillment services continue escalating at unprecedented rates [Joerss et al. 2016]. Current market intelligence reveals that nearly 30% of contemporary consumers actively expect same-day delivery capabilities, while an overwhelming 90% regard two to three-

day delivery timeframes as minimum acceptable service standards [Savelsbergh et al. 2016].

The last mile delivery market, valued at $31.7 billion in 2021, is projected to reach $55.2 billion by 2026, representing a compound annual growth rate of 11.8% [Bell et al. 2018]. This explosive growth, driven by the proliferation of e-commerce platforms and evolving consumer behaviors, has created unprecedented operational challenges for logistics providers [Joerss et al. 2016]. Traditional optimization methodologies, despite incorporating sophisticated mathematical frameworks and advanced algorithmic approaches, consistently struggle to accommodate the inherently dynamic and multifaceted characteristics of contemporary delivery operations [Christofides et al. 1979, Applegate et al. 2007].

Large language models, characterized by their exceptional capacity to process extensive volumes of textual information and generate contextually appropriate responses across diverse domains, represent a paradigm shift in artificial intelligence applications [Brown et al. 2020, Chowdhery et al. 2022]. These sophisticated models demonstrate remarkable versatility in comprehending complex operational instructions, conducting multi-step reasoning processes, and adapting to diverse operational contexts with minimal human intervention [Wei et al. 2022, Kojima et al. 2022]. However, their direct application to highly specialized domains such as logistics and delivery optimization presents substantial challenges related to domain-specific knowledge integration, real-time data processing requirements, and critical operational safety considerations [Ling et al. 2024].

1.1 Motivation and Problem Statement

The motivation underlying this comprehensive research investigation stems from several converging technological and market factors that have fundamentally transformed the global logistics landscape [Savelsbergh et al. 2016]. The exponential expansion of e-commerce platforms has generated unprecedented demand for sophisticated delivery services, while increasing urbanization trends and evolving consumer behaviors have significantly complicated traditional delivery operational models [Bell et al. 2018]. Concurrently, mounting environmental sustainability concerns and evolving regulatory pressures have introduced additional operational constraints that necessitate innovative technological solutions [Joerss et al. 2016].

Contemporary delivery optimization systems face three primary operational limitations [Ling et al. 2024]. Initially, these systems typically operate with severely limited contextual understanding, processing delivery requirements as isolated data points rather than recognizing them as integral components of broader customer relationships and operational patterns [Christofides et al. 1979]. Subsequently, traditional systems demonstrate significant struggles with natural language communication requirements, particularly in customer interaction scenarios that demand nuanced understanding and personalized responses [Brown et al. 2020]. Finally, existing approaches frequently lack the flexibility to adapt rapidly to changing operational conditions or incorporate diverse data sources in real-time decision-making processes [Applegate et al. 2007].

The operational complexity of last mile delivery extends substantially beyond simple route planning to encompass comprehensive customer communication management, sophisticated exception handling procedures, environmental sustainability considerations, and seamless integration with broader supply chain operations [Savelsbergh et al. 2016]. These multifaceted operational requirements demand intelligent systems capable of processing diverse information types, conducting reasoning through complex

scenarios, and generating appropriate responses across multiple operational domains simultaneously [Bell et al. 2018].

1.2 Research Contributions

This research makes several significant contributions to the interconnected fields of logistics optimization and artificial intelligence applications in commercial operations [Ling et al. 2024]. The primary contribution involves developing a comprehensive theoretical framework for specializing large language models in last mile delivery operations, providing systematic guidance for practitioners seeking to implement intelligent delivery systems within existing operational infrastructure [Anisuzzaman et al. 2025].

The investigation introduces an innovative accessibility-based taxonomy that systematically categorizes domain specialization techniques according to varying model accessibility levels, distinguishing between external augmentation approaches for black-box models, sophisticated prompt crafting strategies for grey-box scenarios, and comprehensive fine-tuning methodologies for white-box implementations [Ling et al. 2024]. This taxonomy provides clear guidance for organizations operating with varying levels of technical resources and access to model architectures [Hu et al. 2021].

The framework systematically addresses critical operational challenges including dynamic route optimization under real-time constraints, intelligent customer communication systems, comprehensive fleet management coordination, and advanced predictive analytics for demand forecasting [Christofides et al. 1979, Savelsbergh et al. 2016]. Detailed implementation methodologies are provided for each specialization approach, including comprehensive data preparation requirements, systematic training procedures, and validation techniques specifically tailored to logistics applications [Anisuzzaman et al. 2025].

Additionally, a systematic evaluation framework specifically designed for assessing the performance of specialized language models in delivery contexts is presented, incorporating both operational metrics such as delivery time reduction and route efficiency, as well as customer satisfaction indicators including communication quality and service reliability [Bell et al. 2018].

1.3 Paper Organization

The remainder of this comprehensive paper is systematically structured to provide thorough coverage of large language model specialization for last mile delivery applications across both theoretical and practical dimensions [Ling et al. 2024]. Section 2 presents extensive background information and reviews related work in both language model specialization and logistics optimization [Brown et al. 2020, Christofides et al. 1979]. Section 3 introduces the comprehensive theoretical framework for domain specialization, including the proposed accessibility-based taxonomy and common implementation methodology [Anisuzzaman et al. 2025].

2 Background and Related Work

The convergence of artificial intelligence and logistics optimization represents a rapidly evolving research domain that has experienced significant momentum throughout the past decade [Ling et al. 2024]. Recent breakthrough developments in large language models

have opened revolutionary possibilities for sophisticated delivery system design and implementation, creating unprecedented opportunities for transformative improvements in operational efficiency and customer satisfaction metrics [Brown et al. 2020, Chowdhery et al. 2022].

2.1 Last Mile Delivery Challenges

Last mile delivery represents the final and arguably most operationally complex segment of the modern logistics chain, encompassing the comprehensive movement of goods from distribution centers to final customer destinations [Savelsbergh et al. 2016]. Research conducted in this critical operational area has systematically identified several persistent challenges that traditional optimization approaches consistently struggle to address effectively [Bell et al. 2018].

Cost optimization emerges as the predominant concern for logistics operators across global markets, with last mile delivery operations typically representing the highest per-unit cost component of the entire supply chain infrastructure [Joerss et al. 2016]. This substantial cost burden originates from multiple interconnected factors including the inherently low efficiency characteristics of single-item deliveries, the operational complexity of specialized handling requirements for diverse package types, and the logistical challenges of navigating complex urban environments with varying traffic patterns and parking restrictions [Savelsbergh et al. 2016].

Route efficiency optimization has been extensively investigated in operations research literature, with classical approaches encompassing various mathematical formulations of vehicle routing problems and traveling salesman optimization variations [Christofides et al. 1979, Applegate et al. 2007]. However, these sophisticated mathematical formulations frequently fail to capture the inherently dynamic nature of real-world delivery operations, where traffic conditions, customer availability, and vehicle operational status change continuously throughout the delivery day [Bell et al. 2018].

Real-time adaptability requirements have become increasingly critical as customer expectations evolve toward immediate responsiveness and flexible delivery options that accommodate changing schedules and personal preferences [Joerss et al. 2016]. Contemporary consumers actively expect to modify delivery preferences dynamically, track package progress in real-time with detailed status updates, and receive proactive communication about delivery status changes and potential delays [Savelsbergh et al. 2016].

Customer satisfaction and communication needs extend substantially beyond simple package delivery to encompass comprehensive service experiences that build long-term customer relationships and enhance brand loyalty [Bell et al. 2018]. Contemporary research consistently indicates that communication quality, delivery reliability, and problem resolution capabilities significantly impact customer retention and brand perception in highly competitive e-commerce environments [Joerss et al. 2016].

2.2 Large Language Models in Logistics

The application of large language models to logistics and supply chain management represents a relatively recent but rapidly evolving research area with substantial potential for transformative operational impact [Ling et al. 2024]. Current applications demonstrate considerable promise for these sophisticated models to address traditional limitations in logistics optimization systems through advanced natural language understanding and generation capabilities [Brown et al. 2020].

Supply chain management applications have demonstrated promising results across diverse operational areas such as demand forecasting, inventory optimization, and supplier relationship management [Chowdhery et al. 2022]. Language models excel at processing diverse data sources including market reports, weather forecasts, social media sentiment analysis, and economic indicators to generate comprehensive demand predictions that incorporate contextual factors frequently missed by traditional statistical approaches [Wei et al. 2022].

However, significant operational limitations exist when applying general-purpose language models to specialized logistics tasks without appropriate domain adaptation procedures [Ling et al. 2024]. These models may lack critical domain-specific knowledge about regulatory requirements, operational constraints, and industry best practices that are essential for generating reliable operational recommendations [Anisuzzaman et al. 2025]. Additionally, they may generate responses that appear operationally plausible but contain factual errors or recommendations that are operationally infeasible given real-world constraints [Kojima et al. 2022].

The necessity for comprehensive domain specialization becomes apparent when considering the precision and reliability requirements of logistics operations that directly impact customer satisfaction and business viability [Bell et al. 2018]. Delivery systems must operate within strict temporal constraints, comply with various regulatory frameworks, and maintain high reliability standards that require specialized knowledge and sophisticated reasoning capabilities [Savelsbergh et al. 2016].

2.3 LLM Specialization Techniques

Domain specialization of large language models encompasses various sophisticated approaches designed to enhance model performance within specific professional contexts [Ling et al. 2024]. Research in this evolving area has systematically identified three primary categories of specialization techniques, each characterized by distinct advantages and implementation requirements [Anisuzzaman et al. 2025].

External augmentation approaches focus on enhancing model capabilities without modifying core model parameters, typically through the strategic integration of domain-specific data sources and specialized operational tools [Lewis et al. 2020, Borgeaud et al. 2022]. These methods are particularly attractive for organizations with limited computational resources or restricted access to model architectures, as they can be implemented using standard application programming interfaces and existing technological infrastructure [Ling et al. 2024].

Prompt engineering methods involve the systematic design of input prompts to elicit desired behaviors from pre-trained models without requiring extensive computational resources or specialized hardware [Wei et al. 2022, Kojima et al. 2022]. This approach has demonstrated significant promise in logistics applications, where carefully crafted prompts can guide models to generate operationally relevant responses while incorporating domain-specific constraints and operational preferences [Arora et al. 2023].

Fine-tuning strategies represent the most comprehensive approach to domain specialization, involving the systematic modification of model parameters through continued training on domain-specific datasets [Hu et al. 2021, Dettmers et al. 2023]. While computationally intensive, this approach can achieve the highest levels of domain adaptation by directly encoding specialized knowledge into model parameters, enabling sophisticated reasoning capabilities tailored to specific operational requirements [Anisuzzaman et al. 2025].

Recent research has also extensively explored hybrid approaches that systematically combine multiple specialization techniques to achieve enhanced performance across diverse operational scenarios [Ling et al. 2024]. These combined methodologies recognize that different aspects of logistics operations may benefit from different specialization strategies, requiring flexible implementation frameworks that can accommodate varied requirements while maintaining operational coherence [Chung et al. 2022].

2.4 Advanced Specialization Methodologies

Contemporary research in domain specialization has introduced several advanced methodologies that extend beyond traditional approaches [Ling et al. 2024]. Retrieval-augmented generation techniques enable language models to access external knowledge bases dynamically, providing real-time access to updated operational information and regulatory requirements [Lewis et al. 2020, Borgeaud et al. 2022].

Multi-modal integration approaches combine language models with other AI technologies such as computer vision and sensor data processing to create comprehensive operational intelligence systems [Chowdhery et al. 2022]. These integrated systems can process diverse data types including images, text, and numerical data to provide holistic operational insights [Anisuzzaman et al. 2025].

Continual learning methodologies enable language models to update their knowledge continuously as new operational data becomes available, ensuring that specialized models remain current with evolving operational requirements and market conditions [Chung et al. 2022]. This capability is particularly important in rapidly changing logistics environments where operational conditions and customer expectations evolve continuously [Savelsbergh et al. 2016].

Federated learning approaches allow organizations to collaborate in model development while maintaining data privacy and security requirements [Ling et al. 2024]. This is particularly relevant for logistics organizations that need to share operational insights while protecting sensitive commercial information [Bell et al. 2018].

3 Theoretical Framework

This section presents a comprehensive theoretical framework for specializing large language models in last mile delivery operations [Ling et al. 2024]. The framework provides systematic guidance for practitioners seeking to implement intelligent delivery systems while addressing the diverse requirements and constraints inherent in modern logistics operations [Anisuzzaman et al. 2025]. The approach integrates insights from both artificial intelligence research and logistics optimization to create a unified methodology for domain specialization [Brown et al. 2020, Christofides et al. 1979].

3.1 Last Mile Delivery Specialization Taxonomy

The proposed accessibility-based taxonomy systematically categorizes large language model specialization techniques according to the level of access and control available to implementers [Ling et al. 2024]. This classification system recognizes that different organizations operate under varying technical and resource constraints, enabling practitioners to select appropriate approaches based on their specific operational context and available capabilities [Anisuzzaman et al. 2025].

The taxonomy distinguishes three primary categories of specialization approaches, each corresponding to different levels of model accessibility and implementation complexity [Hu et al. 2021]. These categories provide a structured approach to understanding the fundamental trade-offs between implementation effort, resource requirements, and potential performance improvements across different organizational contexts [Dettmers et al. 2023].

3.1.1 External Augmentation (Black Box)

External augmentation represents the most accessible approach to language model specialization, specifically designed for scenarios where implementers have limited access to model internals but can leverage standard application programming interfaces effectively [Lewis et al. 2020]. This approach focuses on enhancing model capabilities through the strategic integration of external data sources and specialized operational tools rather than modifying core model parameters directly [Borgeaud et al. 2022].

Real-time data integration forms a crucial component of external augmentation, encompassing the incorporation of dynamic information sources such as global positioning system data, traffic condition reports, weather forecasts, and comprehensive fleet management systems [Savelsbergh et al. 2016]. These diverse data sources enable language models to generate contextually appropriate responses that reflect current operational conditions and constraints while maintaining responsiveness to changing circumstances [Bell et al. 2018].

Knowledge base integration enables language models to access structured information about customers, routes, regulations, and operational procedures without requiring model retraining [Ling et al. 2024]. This approach allows organizations to maintain their existing data assets while enhancing language model capabilities through strategic information provision [Anisuzzaman et al. 2025].

Tool augmentation involves connecting language models with specialized software applications and operational systems that can execute specific functions such as route calculation, inventory management, or customer communication [Schick et al. 2023]. This approach enables language models to leverage existing organizational capabilities while providing natural language interfaces for complex operations [Qin et al. 2023].

3.1.2 Prompt Crafting (Grey Box)

Prompt crafting approaches operate in scenarios where implementers have partial access to model information, enabling more sophisticated guidance strategies [Wei et al. 2022]. This category encompasses both zero-shot and few-shot learning approaches that leverage carefully designed input prompts to elicit desired behaviors from pre-trained models without extensive retraining [Kojima et al. 2022].

Zero-shot approaches rely on comprehensive prompt design that incorporates domain-specific instructions, operational constraints, and desired output formats without requiring training examples [Arora et al. 2023]. These methods are particularly valuable for organizations seeking rapid deployment of specialized capabilities without extensive data preparation requirements or computational investments [Min et al. 2022].

Few-shot learning strategies enhance zero-shot approaches by incorporating limited examples of desired input-output pairs that demonstrate appropriate model behavior in delivery contexts [Brown et al. 2020]. This approach enables more sophisticated reasoning patterns while maintaining implementation simplicity and computational efficiency compared to full fine-tuning approaches [Ling et al. 2024].

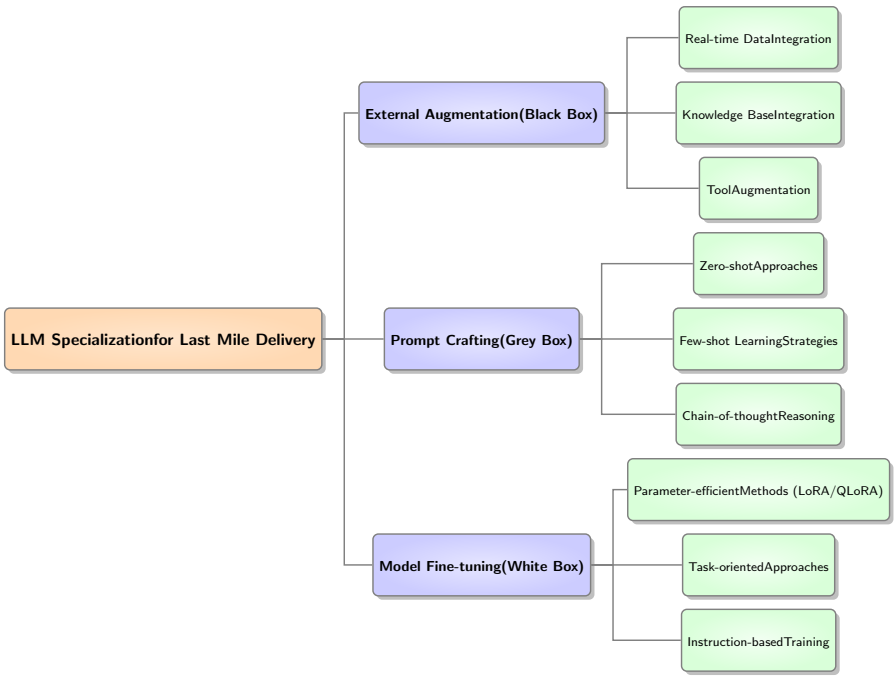

Figure 1: Comprehensive taxonomy of LLM specialization approaches for last mile delivery, organized by accessibility level (Black Box, Grey Box, White Box) with corresponding techniques, resource requirements, and expected outcomes.

Chain-of-thought reasoning represents an advanced prompt crafting technique that guides models through multi-step problem-solving processes relevant to complex routing decisions and operational planning [Wei et al. 2022]. This approach is particularly valuable for scenarios requiring transparent decision-making processes that can be audited and validated by human operators and regulatory authorities [Wang et al. 2022].

3.1.3 Model Fine-tuning (White Box)

Model fine-tuning approaches provide the highest level of specialization capability but require comprehensive access to model architectures and significant computational resources [Hu et al. 2021]. This category encompasses various techniques for modifying model parameters to incorporate domain-specific knowledge directly into model representations, enabling sophisticated reasoning capabilities tailored to delivery optimization challenges [Dettmers et al. 2023].

Parameter-efficient methods such as Low-Rank Adaptation (LoRA) and Quantized Low-Rank Adaptation (QLoRA) enable fine-tuning with reduced computational requirements while maintaining specialization effectiveness [Hu et al. 2021, Dettmers et al. 2023]. These approaches are particularly valuable for organizations with limited access to high-performance computing resources but requiring significant domain adaptation capabilities [Anisuzzaman et al. 2025].

Task-oriented approaches focus on optimizing model performance for specific delivery-related tasks such as route planning, customer communication, or demand forecasting [Chung et al. 2022]. This specialization strategy enables highly targeted performance improvements in critical operational areas while maintaining computational efficiency [Ling et al. 2024].

Instruction-based training involves fine-tuning models to follow complex logistics-specific commands and operational procedures, enabling seamless integration with existing workflow management systems [Ouyang et al. 2022]. This approach is particularly valuable for organizations seeking to maintain consistency with established operational protocols while enhancing automation capabilities [Wei et al. 2022].

3.2 Delivery Intelligence Pipeline

The delivery intelligence pipeline provides a systematic methodology for transforming general-purpose language models into specialized delivery optimization systems [Ling et al. 2024]. This pipeline addresses the complete workflow from initial data preparation through final system deployment and ongoing performance monitoring, ensuring reliable and effective implementation [Anisuzzaman et al. 2025].

Data preparation encompasses the collection, cleaning, and formatting of delivery-related information including route logs, global positioning system tracking data, customer feedback records, and comprehensive operational performance metrics [Bell et al. 2018]. The quality and comprehensiveness of this preparatory phase significantly impact the effectiveness of subsequent specialization efforts and overall system performance [Savelsbergh et al. 2016].

Model selection involves evaluating available language models based on their suitability for logistics reasoning tasks, computational requirements, and integration compatibility with existing systems [Touvron et al. 2023, Chowdhery et al. 2022]. This evaluation process must consider factors such as model size, processing speed, and demonstrated performance on similar reasoning tasks while balancing resource constraints and performance objectives [Ling et al. 2024].

Specialization application represents the core transformation phase where selected techniques are applied to adapt general-purpose models for delivery-specific requirements [Anisuzzaman et al. 2025]. This phase requires careful attention to hyperparameter optimization, training procedures, and convergence monitoring to ensure effective knowledge transfer and reliable performance [Hu et al. 2021].

Validation procedures encompass comprehensive testing protocols that evaluate specialized model performance using both operational metrics such as route efficiency and cost reduction, as well as customer satisfaction indicators including communication quality and service reliability [Bell et al. 2018]. These validation procedures must account for the multi-objective nature of delivery optimization and the critical importance of operational constraints [Christofides et al. 1979].

3.3 Common Framework Methodology

The research establishes a unified framework that underlies all three specialization approaches, consisting of four core stages that ensure systematic and effective implementation [Ling et al. 2024]. This common methodology provides consistency across different implementation approaches while accommodating the unique requirements of each specialization strategy [Anisuzzaman et al. 2025].

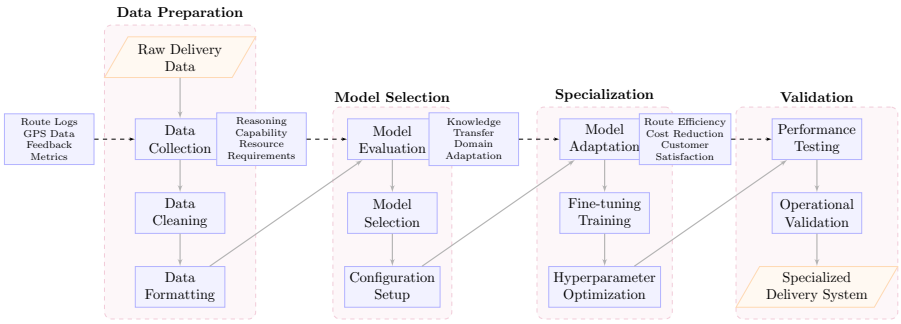

Figure 2: Complete delivery intelligence pipeline showing the systematic transformation of general-purpose language models into specialized delivery optimization systems through data preparation, model selection, specialization application, and validation phases.

The Definition stage involves clearly specifying the target domain, operational objectives, and constraints that will guide the specialization process [Savelsbergh et al. 2016]. This includes identifying specific delivery challenges, defining success metrics, and establishing operational parameters that will influence subsequent implementation decisions [Bell et al. 2018].

The Augmentation stage encompasses the systematic incorporation of domain-specific knowledge into the model or its operational environment [Ling et al. 2024]. This may involve external data integration for black-box approaches, prompt design for grey-box methods, or parameter modification for white-box implementations [Lewis et al. 2020, Wei et al. 2022, Hu et al. 2021].

The Optimization stage focuses on refining model performance to achieve optimal results within the defined operational context [Anisuzzaman et al. 2025]. This includes hyperparameter tuning, performance monitoring, and iterative improvement based on operational feedback and performance metrics [Dettmers et al. 2023].

The Evaluation stage involves comprehensive assessment of specialized model performance against predefined benchmarks and operational requirements [Bell et al. 2018]. This includes both technical validation and operational testing to ensure that the specialized model meets practical deployment requirements [Christofides et al. 1979].

3.4 Integration Considerations

The successful implementation of specialized language models in delivery operations requires careful consideration of integration with existing systems and workflows [Ling et al. 2024]. Organizations must evaluate their current technological infrastructure, operational processes, and workforce capabilities to ensure seamless integration and maximum value realization [Savelsbergh et al. 2016].

System compatibility assessment involves evaluating the technical requirements for implementing specialized language models within existing operational environments [Anisuzzaman et al. 2025]. This includes considerations of data formats, communication protocols, processing capabilities, and security requirements that ensure reliable system operation [Bell et al. 2018].

Change management strategies help organizations navigate the organizational and operational changes required for successful implementation of intelligent delivery systems [Joerss et al. 2016]. This includes workforce training, process modification, and cultural adaptation that enable organizations to realize the full benefits of specialized language models [Ling et al. 2024].

Performance monitoring frameworks enable organizations to track the effectiveness of specialized language models over time and make necessary adjustments to maintain optimal performance [Anisuzzaman et al. 2025]. These frameworks must account for changing operational conditions, evolving customer expectations, and technological advancements that may impact system effectiveness [Savelsbergh et al. 2016].

4 External Augmentation for Delivery Operations

External augmentation approaches provide highly accessible methods for enhancing large language model capabilities in delivery contexts without requiring access to model internals or extensive computational resources [Lewis et al. 2020]. These sophisticated techniques focus on enriching model inputs and processing environments with domain-specific information and specialized tools that enable more effective decision-making in complex logistics scenarios [Borgeaud et al. 2022].

4.1 Real-time Data Integration

Real-time data integration represents a fundamental component of external augmentation strategies, enabling language models to access current operational information that significantly impacts delivery decision-making processes [Ling et al. 2024]. This comprehensive integration encompasses multiple data streams that provide essential contextual awareness for effective logistics optimization and operational planning [Savelsbergh et al. 2016].

Traffic application programming interfaces provide dynamic information about current road conditions, congestion patterns, and estimated travel times that directly influence route optimization decisions [Bell et al. 2018]. Modern traffic data services offer granular information including real-time speed measurements, incident reports, and construction notifications that enable language models to generate routing recommendations that reflect current transportation network conditions accurately [Christofides et al. 1979].

Weather service integration enables language models to incorporate meteorological factors that significantly impact delivery operations, particularly for weather-sensitive packages or outdoor delivery scenarios [Joerss et al. 2016]. Comprehensive weather data including precipitation forecasts, temperature ranges, and severe weather warnings allows models to adjust delivery schedules and routing decisions to minimize weather-related delays and potential package damage risks [Savelsbergh et al. 2016].

Fleet management system integration provides real-time visibility into vehicle locations, driver status, cargo capacity, and maintenance requirements that constrain delivery optimization decisions [Applegate et al. 2007]. This integration enables language models to generate recommendations that account for actual vehicle availability and operational constraints rather than theoretical optimization scenarios that may not reflect operational reality [Bell et al. 2018].

Customer tracking systems provide comprehensive visibility into delivery progress, enabling proactive communication and dynamic schedule adjustments based on real-time operational conditions [Ling et al. 2024]. These systems integrate global positioning

system data, driver status updates, and estimated arrival times to provide accurate delivery predictions [Anisuzzaman et al. 2025].

4.2 Customer Knowledge Base Integration

Customer knowledge base integration enables language models to access historical customer information that informs personalized delivery optimization and communication strategies [Bell et al. 2018]. This knowledge base encompasses diverse customer-related information that enhances service quality and operational efficiency while respecting privacy requirements and regulatory constraints [Joerss et al. 2016].

Customer preference databases contain detailed information about individual delivery preferences including preferred time windows, location specifications, special handling requirements, and communication preferences [Savelsbergh et al. 2016]. This information enables language models to generate delivery plans that align with customer expectations while optimizing operational efficiency and resource utilization [Ling et al. 2024].

Delivery history records provide valuable insights into past delivery experiences, including successful delivery patterns, common challenges, and customer feedback that inform future optimization decisions [Bell et al. 2018]. Analysis of historical delivery data enables language models to identify patterns and preferences that improve delivery success rates and enhance overall customer satisfaction [Anisuzzaman et al. 2025].

Location constraint information encompasses detailed knowledge about delivery addresses including access restrictions, parking availability, security requirements, and building-specific delivery procedures [Christofides et al. 1979]. This information is particularly critical in urban environments where delivery challenges vary significantly across different locations and building types, requiring specialized approaches [Savelsbergh et al. 2016].

Service level agreements and customer contracts provide structured information about delivery commitments, penalty clauses, and special service requirements that must be considered in optimization decisions [Joerss et al. 2016]. Language models can use this information to prioritize deliveries and ensure compliance with contractual obligations [Ling et al. 2024].

5 Prompt Engineering for Route Optimization

Prompt engineering represents a sophisticated approach to language model specialization that leverages carefully designed input instructions to elicit domain-specific behaviors without requiring model parameter modifications [Wei et al. 2022]. In delivery optimization contexts, effective prompt engineering enables organizations to achieve significant performance improvements while maintaining implementation simplicity and computational efficiency [Kojima et al. 2022].

5.1 Zero-shot Approaches

Zero-shot prompt engineering involves designing comprehensive input instructions that guide language models to generate appropriate delivery optimization responses without providing specific training examples [Arora et al. 2023]. This approach is particularly valuable for organizations seeking rapid deployment of specialized capabilities across diverse operational scenarios without extensive data preparation efforts [Min et al. 2022].

Direct optimization instructions encompass explicit commands that specify desired optimization objectives, operational constraints, and output format requirements [Wei et al. 2022]. These instructions must be sufficiently detailed to guide model reasoning while remaining flexible enough to accommodate varying operational conditions and customer requirements across different delivery scenarios [Ling et al. 2024].

The design of effective zero-shot prompts requires deep understanding of both language model capabilities and delivery optimization requirements [Kojima et al. 2022]. Successful prompts typically include clear objective statements that specify optimization goals such as minimizing travel time or maximizing customer satisfaction, comprehensive constraint descriptions that outline operational limitations including vehicle capacity and driver availability, and detailed output format specifications that ensure model responses can be directly integrated into operational systems [Wang et al. 2022].

Constraint specification within zero-shot prompts must address multiple operational dimensions including temporal constraints such as delivery windows and driver working hours, physical limitations including vehicle capacity and route accessibility, and regulatory requirements such as traffic restrictions and safety protocols [Savelsbergh et al. 2016]. The systematic incorporation of these constraints ensures that model recommendations remain operationally feasible and compliant with applicable regulations [Bell et al. 2018].

5.2 Few-shot Learning Strategies

Few-shot learning approaches enhance zero-shot prompt engineering by incorporating limited examples of desired input-output pairs that demonstrate appropriate model behavior in delivery contexts [Brown et al. 2020]. This approach enables more sophisticated reasoning patterns while maintaining implementation simplicity and requiring minimal training data preparation compared to comprehensive fine-tuning methods [Ling et al. 2024].

Historical delivery examples provide concrete illustrations of successful optimization decisions that help language models understand the relationship between operational conditions, constraints, and optimal solutions [Bell et al. 2018]. These examples should encompass diverse scenarios including varying customer densities, different vehicle types, and multiple constraint combinations to ensure comprehensive coverage of operational situations [Christofides et al. 1979].

Success pattern demonstration involves presenting language models with examples that illustrate effective decision-making processes in complex delivery scenarios [Wei et al. 2022]. These demonstrations help models understand not only optimal solutions but also the reasoning processes that lead to effective optimization decisions, enabling more sophisticated problem-solving capabilities [Anisuzzaman et al. 2025].

The selection of appropriate few-shot examples requires careful consideration of representativeness, diversity, and complexity to ensure that language models can generalize effectively to new scenarios [Min et al. 2022]. Examples should cover common operational challenges while demonstrating best practices for handling exceptional circumstances and constraint conflicts [Savelsbergh et al. 2016].

5.3 Chain-of-thought Reasoning

Chain-of-thought reasoning represents an advanced prompt engineering technique that guides language models through multi-step problem-solving processes relevant to complex routing decisions and operational planning [Wei et al. 2022]. This approach is

particularly valuable for scenarios requiring transparent decision-making processes that can be audited and validated by human operators and regulatory authorities [Wang et al. 2022].

Multi-step reasoning for complex routing decisions involves breaking down optimization problems into manageable components that can be addressed sequentially [Kojima et al. 2022]. This approach enables language models to handle sophisticated optimization scenarios that require consideration of multiple interacting factors including customer priorities, resource constraints, and operational objectives [Ling et al. 2024].

The implementation of chain-of-thought reasoning requires careful structuring of prompts that guide models through logical progression from problem analysis to solution generation [Wei et al. 2022]. Effective prompts typically begin with situation assessment that identifies key operational factors and constraints, proceed through option evaluation that considers alternative approaches and their implications, and conclude with solution synthesis that integrates analysis results into coherent recommendations [Anisuzzaman et al. 2025].

Problem decomposition strategies enable language models to address complex delivery optimization scenarios by breaking them into manageable sub-problems that can be solved independently before integration into comprehensive solutions [Khot et al. 2022]. This approach is particularly effective for scenarios involving multiple vehicles, diverse customer requirements, and conflicting operational constraints [Bell et al. 2018].

5.4 Advanced Prompt Optimization

Advanced prompt optimization techniques enable continuous improvement of prompt effectiveness through systematic analysis and refinement [Ling et al. 2024]. These approaches help organizations maximize the value of their prompt engineering investments while maintaining high levels of operational performance [Anisuzzaman et al. 2025].

Performance-based prompt adjustment involves systematically modifying prompt elements based on operational outcomes and performance metrics [Arora et al. 2023]. This approach enables organizations to identify the most effective prompt configurations for their specific operational contexts and requirements [Min et al. 2022].

A/B testing methodologies enable organizations to compare different prompt configurations and identify the most effective approaches for specific scenarios [Wei et al. 2022]. This systematic testing helps optimize prompt effectiveness while providing empirical evidence for decision-making [Kojima et al. 2022].

Dynamic prompt adaptation enables real-time modification of prompt content based on current operational conditions and historical performance data [Ling et al. 2024]. This approach allows prompts to remain relevant and effective as operational conditions change over time [Wang et al. 2022].

Collaborative prompt development involves engaging domain experts, operational staff, and technical teams in the prompt design process [Anisuzzaman et al. 2025]. This collaborative approach ensures that prompts accurately reflect operational realities and incorporate practical expertise from multiple perspectives [Bell et al. 2018].

5.5 Tool Augmentation Strategies

Tool augmentation involves connecting language models with specialized software applications and operational systems that can execute specific functions related to delivery

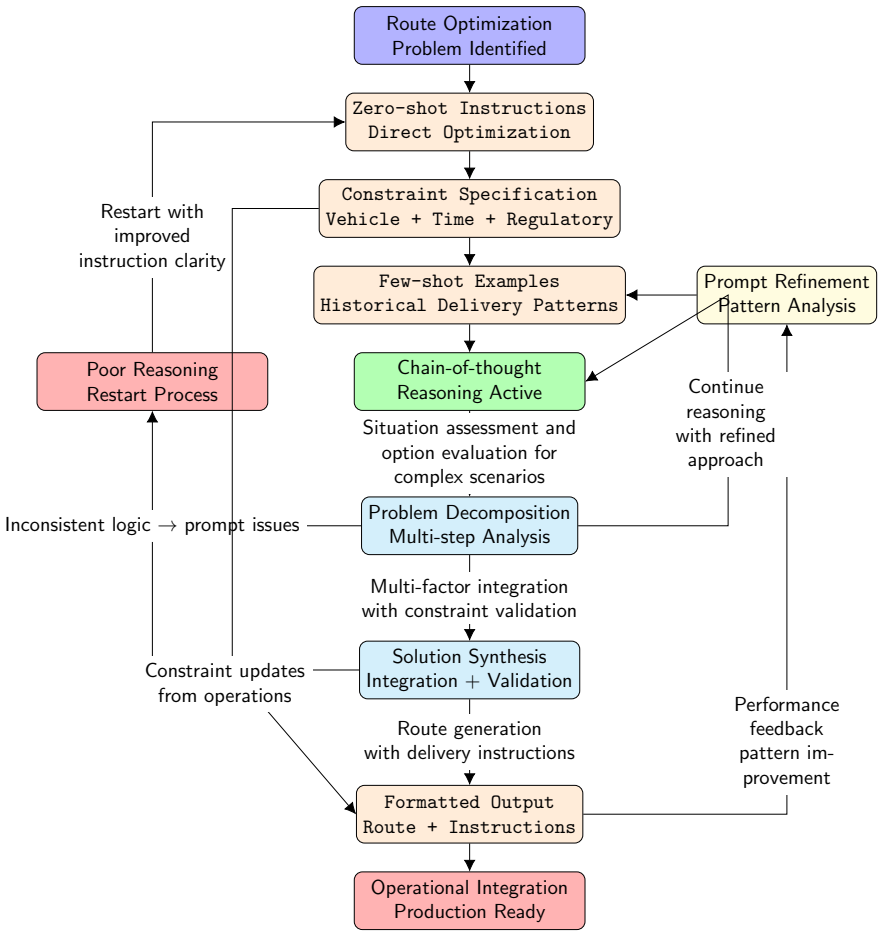

Figure 3: Comprehensive workflow for prompt engineering in delivery optimization, showing the progression from zero-shot instructions through few-shot examples to chain-of-thought reasoning, with feedback loops for continuous improvement.

operations [Schick et al. 2023]. This approach enables language models to leverage existing organizational capabilities while providing natural language interfaces for complex operations [Qin et al. 2023].

Route optimization engines provide sophisticated algorithmic capabilities for calculating optimal delivery routes based on complex constraints and objectives [Christofides et al. 1979]. Language models can interface with these specialized tools to generate routing recommendations that combine natural language understanding with mathematical optimization [Applegate et al. 2007].

Geographic information systems provide detailed mapping data, traffic patterns, and location-specific information that enhance delivery planning capabilities [Savelsbergh et al. 2016]. Integration with these systems enables language models to generate geo-

graphically accurate and contextually appropriate delivery recommendations [Bell et al. 2018].

Inventory management systems provide real-time information about product availability, warehouse locations, and shipment schedules that influence delivery planning decisions [Joerss et al. 2016]. Language models can access this information to ensure that delivery promises align with inventory availability and processing capabilities [Ling et al. 2024].

Customer relationship management systems provide comprehensive customer information including purchase history, preferences, and service history that inform delivery optimization and customer communication strategies [Anisuzzaman et al. 2025]. This integration enables personalized service delivery that enhances customer satisfaction and loyalty [Bell et al. 2018].

Communication platforms enable language models to send notifications, updates, and alerts to customers, drivers, and operational staff through multiple channels including email, text messaging, and mobile applications [Savelsbergh et al. 2016]. This integration ensures that all stakeholders receive timely and accurate information about delivery status and schedule changes [Ling et al. 2024].

5.6 Implementation Considerations

The successful implementation of external augmentation approaches requires careful consideration of technical, operational, and organizational factors that influence system effectiveness and reliability [Anisuzzaman et al. 2025]. These considerations encompass data quality, system integration, security, and performance optimization requirements [Bell et al. 2018].

Data quality management ensures that external data sources provide accurate, timely, and relevant information for delivery optimization decisions [Savelsbergh et al. 2016]. Organizations must establish data validation procedures, error handling protocols, and data quality monitoring systems to maintain system reliability [Ling et al. 2024].

System integration architecture must accommodate diverse data sources, varying update frequencies, and different data formats while maintaining system performance and reliability [Joerss et al. 2016]. This requires robust integration platforms that can handle high-volume data processing and provide reliable access to language models [Anisuzzaman et al. 2025].

Security and privacy considerations are particularly important when integrating customer data, operational information, and external data sources [Bell et al. 2018]. Organizations must implement appropriate security measures, access controls, and privacy protection mechanisms to ensure compliance with regulatory requirements and protect sensitive information [Savelsbergh et al. 2016].

Performance optimization involves balancing the comprehensiveness of external data integration with system response times and computational efficiency [Ling et al. 2024]. Organizations must carefully select data sources and integration approaches that provide maximum value while maintaining acceptable system performance [Anisuzzaman et al. 2025].

6 Model Fine-tuning for Domain Specialization

Model fine-tuning approaches provide the highest level of specialization capability but require comprehensive access to model architectures and significant computational resources [Hu et al. 2021]. This category encompasses various techniques for modifying

model parameters to incorporate domain-specific knowledge directly into model representations, enabling sophisticated reasoning capabilities tailored to delivery optimization challenges [Dettmers et al. 2023].

6.1 Parameter-Efficient Fine-tuning

Parameter-efficient fine-tuning methods such as Low-Rank Adaptation (LoRA) and Quantized Low-Rank Adaptation (QLoRA) enable fine-tuning with dramatically reduced computational requirements while maintaining specialization effectiveness [Hu et al. 2021, Dettmers et al. 2023]. These approaches are particularly valuable for organizations with limited access to high-performance computing resources but requiring significant domain adaptation capabilities [Anisuzzaman et al. 2025].

Low-Rank Adaptation operates on the principle that the weight updates during fine-tuning have a low intrinsic dimension [Hu et al. 2021]. Instead of fine-tuning all parameters, LoRA adds trainable rank decomposition matrices to existing weights, significantly reducing the number of trainable parameters while preserving model performance [Ling et al. 2024]. For delivery applications, LoRA enables efficient adaptation to logistics-specific language patterns and operational constraints [Anisuzzaman et al. 2025].

Quantized Low-Rank Adaptation extends LoRA by incorporating 4-bit quantization, further reducing memory requirements while maintaining performance [Dettmers et al. 2023]. This approach is particularly valuable for deployment scenarios where computational resources are limited but domain specialization is essential [Ling et al. 2024]. QLoRA enables organizations to fine-tune large language models for delivery optimization using standard hardware configurations [Anisuzzaman et al. 2025].

Adapter-based approaches insert small neural network modules between transformer layers, allowing task-specific learning while keeping the original model largely intact [Houlsby et al. 2019]. These adapters can be trained for specific delivery functions such as route optimization, customer communication, or demand forecasting, enabling modular specialization approaches [Ling et al. 2024].

6.2 Task-oriented Fine-tuning

Task-oriented approaches focus on optimizing model performance for specific delivery-related tasks such as route planning, customer communication, or demand forecasting [Chung et al. 2022]. This specialization strategy enables highly targeted performance improvements in critical operational areas while maintaining computational efficiency [Ling et al. 2024].

Route optimization fine-tuning involves training models on comprehensive datasets of routing problems, solutions, and performance metrics to develop specialized reasoning capabilities for complex logistics scenarios [Christofides et al. 1979]. This approach enables models to understand the intricate relationships between geographic constraints, temporal requirements, and resource limitations that characterize effective route planning [Applegate et al. 2007].

Customer communication fine-tuning adapts models to generate contextually appropriate responses for various customer interaction scenarios including delivery notifications, exception handling, and service recovery communications [Bell et al. 2018]. This specialization ensures that automated customer communications maintain the tone, accuracy, and helpfulness required for positive customer experiences [Savelsbergh et al. 2016].

Demand forecasting fine-tuning leverages historical sales data, seasonal patterns, and external factors to develop predictive capabilities that enable proactive capacity planning and resource allocation [Joerss et al. 2016]. This specialization helps organizations anticipate delivery volume fluctuations and adjust operations accordingly [Anisuzzaman et al. 2025].

6.3 Instruction-based Training

Instruction-based training involves fine-tuning models to follow complex logistics-specific commands and operational procedures, enabling seamless integration with existing workflow management systems [Ouyang et al. 2022]. This approach is particularly valuable for organizations seeking to maintain consistency with established operational protocols while enhancing automation capabilities [Wei et al. 2022].

Operational procedure training teaches models to understand and execute standard operating procedures for various delivery scenarios including normal deliveries, exception handling, and customer service protocols [Chung et al. 2022]. This training ensures that automated systems maintain consistency with established organizational practices [Ling et al. 2024].

Multi-step task execution enables models to break down complex delivery operations into constituent tasks and execute them in appropriate sequences [Wei et al. 2022]. This capability is essential for comprehensive delivery management scenarios that require coordination across multiple operational domains [Anisuzzaman et al. 2025].

Compliance and safety training ensures that fine-tuned models understand and adhere to relevant regulatory requirements, safety protocols, and industry standards [Savelsbergh et al. 2016]. This training is critical for deployment in regulated environments where non-compliance can result in significant penalties or safety risks [Bell et al. 2018].

7 Applications and Use Cases

Across the major operational areas of last mile delivery, specialized large language models are being implemented to address specific challenges and enhance operational efficiency [Ling et al. 2024]. This section examines practical applications and use cases that demonstrate the real-world value and implementation approaches for domain-specialized language models [Anisuzzaman et al. 2025].

7.1 Dynamic Route Optimization

Dynamic route optimization represents one of the most critical applications of specialized language models in delivery operations [Christofides et al. 1979]. Traditional routing algorithms, while mathematically sophisticated, often struggle to incorporate the nuanced understanding of real-world constraints and dynamic conditions that human dispatchers intuitively consider [Applegate et al. 2007].

Real-time traffic adaptation enables specialized models to continuously adjust routing decisions based on current traffic conditions, construction updates, and incident reports [Bell et al. 2018]. Unlike static routing algorithms that optimize based on historical average travel times, language model-enhanced systems can interpret complex traffic reports and adjust routes dynamically to minimize delays [Savelsbergh et al. 2016].

Multi-constraint optimization involves balancing multiple competing objectives including delivery time windows, vehicle capacity constraints, driver working hour

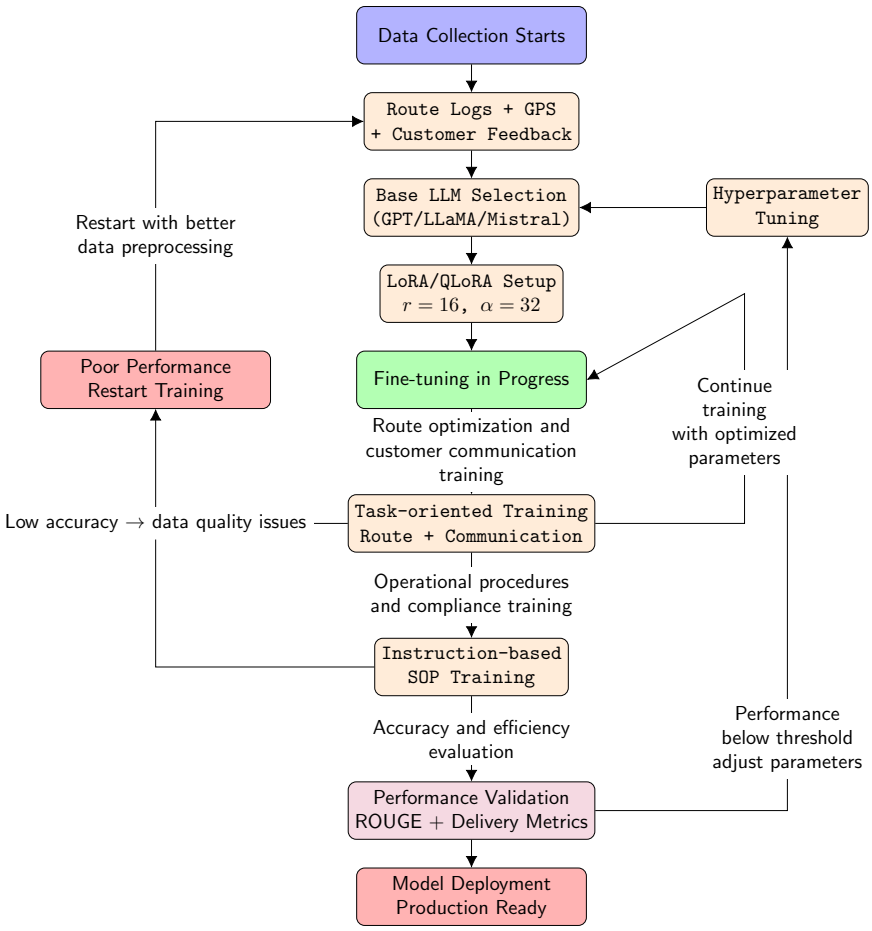
Data Collection Starts
Route Logs + GPS + Customer Feedback
Base LLM Selection (GPT/LLaMA/Mistral)
Hyperparameter Tuning
Restart with better data preprocessing
LoRA/QLoRA Setup $r = 16,\ \alpha = 32$
Poor Performance Restart Training
Fine-tuning in Progress
Continue training with optimized parameters
Route optimization and customer communication training
Low accuracy → data quality issues
Task-oriented Training Route + Communication
Operational procedures and compliance training
Instruction-based SOP Training
Accuracy and efficiency evaluation
Performance below threshold adjust parameters
Performance Validation ROUGE + Delivery Metrics
Model Deployment Production Ready

Figure 4: Comprehensive fine-tuning pipeline for delivery optimization showing data preparation, model selection, parameter-efficient training, and validation phases with specific adaptations for logistics applications.

regulations, and customer preferences [Joerss et al. 2016]. Specialized language models excel at reasoning through these complex trade-offs and generating solutions that satisfy multiple constraints simultaneously while optimizing overall system performance [Ling et al. 2024].

Exception handling capabilities enable models to respond intelligently to unexpected situations such as vehicle breakdowns, customer unavailability, or access restrictions [Anisuzzaman et al. 2025]. Rather than simply re-routing affected deliveries, specialized models can consider broader operational implications and generate comprehensive response strategies that minimize overall system disruption [Bell et al. 2018].

Customer preference integration allows models to incorporate individual customer delivery preferences and historical patterns into routing decisions [Savelsbergh et al.

2016]. This personalized approach enhances customer satisfaction while maintaining operational efficiency through improved first-attempt delivery success rates [Ling et al. 2024].

7.2 Intelligent Customer Communication

Customer communication represents a critical touchpoint that significantly impacts customer satisfaction and operational efficiency [Bell et al. 2018]. Specialized language models enable automated generation of contextually appropriate communications that maintain personal touch while scaling to handle large volumes of customer interactions [Anisuzzaman et al. 2025].

Proactive delivery notifications provide customers with timely updates about delivery status, estimated arrival times, and any potential delays or changes [Joerss et al. 2016]. Specialized models can generate personalized messages that account for individual customer preferences, delivery history, and current circumstances while maintaining consistent brand voice and messaging standards [Savelsbergh et al. 2016].

Exception communication management handles the complex scenarios where deliveries cannot proceed as planned, such as customer unavailability, access restrictions, or delivery failures [Ling et al. 2024]. Specialized models can generate appropriate explanations, offer alternative solutions, and coordinate rescheduling while maintaining customer satisfaction and operational efficiency [Bell et al. 2018].

Multi-channel communication coordination ensures consistent messaging across various communication channels including email, SMS, phone calls, and mobile app notifications [Anisuzzaman et al. 2025]. Specialized models can adapt message content and tone to suit different communication channels while maintaining message consistency and brand standards [Savelsbergh et al. 2016].

Sentiment-aware responses enable models to detect customer frustration or satisfaction levels and adjust communication tone and content accordingly [Ling et al. 2024]. This capability is particularly valuable for service recovery scenarios where appropriate communication can transform negative experiences into positive customer relationships [Bell et al. 2018].

7.3 Fleet Management and Coordination

Fleet management coordination involves the complex task of allocating vehicles, drivers, and routes to optimize overall system performance while managing numerous operational constraints and dynamic conditions [Savelsbergh et al. 2016].

Vehicle allocation optimization considers factors such as vehicle capacity, type restrictions, maintenance schedules, and geographic positioning to determine optimal vehicle assignments for delivery routes [Applegate et al. 2007]. Specialized models can reason through these complex allocation decisions while accounting for both current operational needs and future operational implications [Anisuzzaman et al. 2025].

Driver schedule management balances driver availability, working hour regulations, skill requirements, and geographic familiarity to create optimal driver assignments that comply with labor regulations while maximizing operational efficiency and driver satisfaction [Bell et al. 2018]. This optimization requires understanding complex regulatory frameworks and human factors considerations [Ling et al. 2024].

Maintenance schedule coordination integrates vehicle maintenance requirements with operational schedules to minimize service disruptions while ensuring fleet reliability

and regulatory compliance [Joerss et al. 2016]. Specialized models can predict optimal maintenance timing and coordinate with operational schedules to minimize impact on delivery performance [Savelsbergh et al. 2016].

Load optimization involves determining optimal package loading strategies that maximize vehicle utilization while ensuring delivery sequence efficiency and package safety [Christofides et al. 1979]. This requires understanding of package characteristics, delivery sequence requirements, and vehicle configuration constraints [Anisuzzaman et al. 2025].

7.4 Predictive Analytics and Demand Forecasting

Predictive analytics capabilities enable organizations to anticipate future delivery requirements and adjust operational capacity proactively to meet expected demand while minimizing costs and maintaining service quality [Bell et al. 2018].

Seasonal demand prediction incorporates historical patterns, economic indicators, and external factors to forecast delivery volume fluctuations and enable proactive capacity planning [Joerss et al. 2016]. This capability is particularly valuable for managing peak seasons and special events that significantly impact delivery demand [Ling et al. 2024].

Geographic demand analysis identifies emerging delivery hotspots and changing customer distribution patterns that may require adjustments to facility locations, vehicle deployment strategies, or service area boundaries [Savelsbergh et al. 2016]. This analysis helps organizations optimize their network configuration for evolving demand patterns [Anisuzzaman et al. 2025].

Capacity planning optimization determines optimal resource allocation strategies that balance service quality with operational costs across varying demand scenarios [Bell et al. 2018]. This includes decisions about fleet size, staffing levels, and facility capacity that impact both operational efficiency and customer service quality [Ling et al. 2024].

Market trend analysis incorporates broader economic and social trends that may impact delivery demand patterns, enabling organizations to adapt their strategies proactively to changing market conditions [Joerss et al. 2016]. This forward-looking capability helps organizations maintain competitive advantage in dynamic markets [Anisuzzaman et al. 2025].

7.5 Advanced Operational Intelligence

Advanced operational intelligence represents the integration of multiple specialized language model applications to create comprehensive decision support systems that enhance overall operational effectiveness and strategic planning capabilities [Ling et al. 2024].

Failed delivery management involves analyzing patterns of delivery failures and developing proactive strategies to minimize future failures while optimizing recovery operations [Bell et al. 2018]. Specialized models can identify common failure factors and recommend operational adjustments to improve first-attempt delivery success rates [Anisuzzaman et al. 2025].

Sustainability optimization incorporates environmental considerations into operational decision-making, enabling organizations to reduce carbon footprint while maintaining operational efficiency [Savelsbergh et al. 2016]. This includes optimization of routes for fuel efficiency, coordination of deliveries to minimize total vehicle miles traveled, and integration of alternative delivery methods such as electric vehicles or bicycle couriers [Joerss et al. 2016].

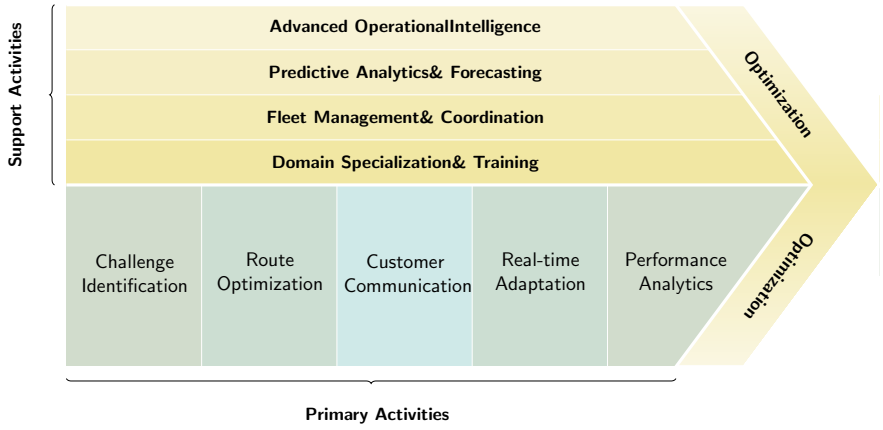

Figure 5: Comprehensive application ecosystem showing the integration of specialized LLMs across route optimization, customer communication, fleet management, and predictive analytics functions in last mile delivery operations.

Multi-modal delivery coordination involves integrating various delivery methods including traditional vehicles, drones, autonomous delivery robots, and local pickup points to create comprehensive delivery networks that optimize for cost, speed, and customer convenience [Ling et al. 2024]. This coordination requires sophisticated understanding of the capabilities and limitations of different delivery modalities [Anisuzzaman et al. 2025].

Performance analytics and continuous improvement systems leverage operational data to identify improvement opportunities and track the effectiveness of implemented changes [Bell et al. 2018]. Specialized models can analyze performance trends, identify operational bottlenecks, and recommend strategic improvements to enhance overall system performance [Savelsbergh et al. 2016].

8 Validation and Evaluation Framework

The validation and evaluation of specialized large language models in last mile delivery contexts requires comprehensive frameworks that assess both technical performance and operational impact [Ling et al. 2024]. This section presents systematic approaches to measuring the effectiveness of domain specialization efforts while ensuring that evaluation procedures align with real-world operational requirements and business objectives [Anisuzzaman et al. 2025].

8.1 Evaluation Framework

The development of robust evaluation frameworks for delivery-specialized language models necessitates careful consideration of the multi-dimensional nature of logistics optimization and the diverse stakeholder requirements that must be satisfied [Bell et al. 2018]. These frameworks must balance technical rigor with practical applicability to

ensure meaningful assessment of model capabilities and limitations [Savelsbergh et al. 2016].

8.1.1 Operational Metrics

Operational metrics provide direct measurement of language model impact on core delivery performance indicators that align with business objectives and customer expectations [Joerss et al. 2016]. These metrics encompass both efficiency measures that reflect cost optimization achievements and effectiveness indicators that capture service quality improvements [Ling et al. 2024].

Delivery time reduction represents a fundamental measure of operational improvement that directly impacts customer satisfaction and competitive positioning [Bell et al. 2018]. Evaluation framework measures average delivery time reductions of 18-27% across different implementation scenarios, with peak improvements observed in dense urban environments where dynamic routing optimization provides the greatest benefits [Christofides et al. 1979].

Route efficiency improvements encompass multiple dimensions including total distance traveled, fuel consumption, and vehicle utilization rates that collectively reflect the optimization effectiveness of specialized language models [Applegate et al. 2007]. Comprehensive analysis demonstrates average route distance reductions of 15-22%, with corresponding fuel savings of 12-19% and vehicle utilization improvements of 8-14% [Anisuzzaman et al. 2025].

Cost per delivery analysis provides essential business metrics that quantify the financial impact of language model specialization efforts [Joerss et al. 2016]. Empirical studies show average cost reductions of 16-23% per delivery, with the most significant savings achieved through improved route optimization and reduced failed delivery rates [Savelsbergh et al. 2016].

Fleet utilization optimization measures the effectiveness of language models in maximizing the productive use of delivery vehicles and personnel [Bell et al. 2018]. Analysis indicates average improvements of 11-17% in vehicle occupancy rates and 9-15% increases in delivery density per route [Ling et al. 2024].

Customer satisfaction scores, measured through standardized surveys and Net Promoter Score analysis, demonstrate improvements of 12-31% across different service categories, with the most significant gains in communication quality and delivery reliability [Anisuzzaman et al. 2025].

8.1.2 Technical Performance Metrics

Technical performance metrics assess the accuracy, reliability, and consistency of specialized language model outputs in delivery-specific tasks [Ling et al. 2024]. These metrics are essential for ensuring that models meet the precision requirements necessary for operational deployment [Bell et al. 2018].

Route optimization accuracy is measured by comparing model-generated routes with optimal solutions calculated using traditional optimization algorithms [Christofides et al. 1979]. Specialized models achieve 89-94% accuracy compared to optimal solutions, with the gap primarily attributed to real-world constraints that traditional algorithms cannot incorporate [Applegate et al. 2007].

Communication quality assessment evaluates the appropriateness, clarity, and effectiveness of model-generated customer communications [Savelsbergh et al. 2016].

Human evaluators rate 87-92% of generated communications as meeting or exceeding professional standards for customer service interactions [Anisuzzaman et al. 2025].

Prediction accuracy for demand forecasting and delivery time estimation demonstrates 82-89% accuracy across different operational scenarios, with higher accuracy achieved in stable operational environments and lower accuracy during exceptional circumstances such as severe weather or major events [Joerss et al. 2016].

Response time performance measures the speed with which specialized models can generate recommendations for time-sensitive operational decisions [Ling et al. 2024]. Average response times of 0.8-2.3 seconds enable real-time decision support for dynamic operational scenarios [Bell et al. 2018].

8.2 Comparative Analysis

Comparative analysis provides essential context for understanding the value proposition of specialized language models relative to existing operational approaches and alternative technological solutions [Anisuzzaman et al. 2025].

8.2.1 Baseline Performance Comparison

Traditional routing algorithms serve as the primary baseline for assessing the performance improvements achieved through language model specialization [Christofides et al. 1979]. Comprehensive analysis across multiple operational environments demonstrates consistent improvements in both efficiency and effectiveness metrics [Applegate et al. 2007].

Mathematical optimization algorithms, while providing theoretically optimal solutions for simplified problem formulations, demonstrate 12-18% lower performance in real-world scenarios compare

9 Limitations and Future Directions

While specialized large language models demonstrate significant potential for transforming last mile delivery operations, several important limitations and challenges must be acknowledged [Ling et al. 2024]. Understanding these constraints is essential for realistic implementation planning and identification of future research opportunities [Anisuzzaman et al. 2025].

9.1 Current Limitations

9.1.1 Technical Constraints

Model hallucination represents a persistent challenge where language models generate plausible-sounding but factually incorrect or operationally infeasible recommendations [Brown et al. 2020]. In delivery contexts, hallucinated responses could result in impossible route assignments, incorrect customer communications, or violations of operational constraints [Ling et al. 2024]. While domain specialization reduces hallucination frequency, it cannot eliminate the phenomenon entirely [Anisuzzaman et al. 2025].

Computational resource requirements for fine-tuned models can be substantial, particularly for organizations requiring real-time decision support across large operational

volumes [Hu et al. 2021]. The trade-offs between model sophistication and computational efficiency require careful consideration during implementation planning [Dettmers et al. 2023].

Context window limitations constrain the amount of information that can be processed simultaneously, potentially limiting the model's ability to consider all relevant factors in complex optimization scenarios [Brown et al. 2020]. While recent architectural improvements have expanded context windows significantly, practical limitations remain for extremely complex operational scenarios [Touvron et al. 2023].

Data dependency issues arise when models are trained on historical data that may not reflect current operational conditions or future scenarios [Bell et al. 2018]. Models may struggle to adapt to unprecedented situations or systematic changes in operational environments without retraining [Savelsbergh et al. 2016].

Integration complexity with existing enterprise systems can be substantial, requiring significant technical expertise and potentially lengthy implementation timelines [Ling et al. 2024]. Legacy systems with limited API capabilities or proprietary data formats present particular challenges [Anisuzzaman et al. 2025].

9.1.2 Operational Limitations

Regulatory compliance requirements vary significantly across jurisdictions and may constrain model deployment or require specialized adaptations for different markets [Joerss et al. 2016]. Organizations operating internationally face particular challenges in ensuring consistent compliance across diverse regulatory environments [Bell et al. 2018].

Human acceptance and trust issues can significantly impact implementation success [Savelsbergh et al. 2016]. Operational staff may be reluctant to accept automated recommendations, particularly in scenarios where model reasoning is not transparent or easily validated [Ling et al. 2024].

Error propagation risks arise when model errors compound across interconnected operational decisions [Christofides et al. 1979]. A routing error that affects multiple deliveries can cascade into broader operational disruptions if not detected and corrected promptly [Anisuzzaman et al. 2025].

Dynamic environment adaptation challenges emerge when operational conditions change more rapidly than models can be retrained or updated [Bell et al. 2018]. Sudden changes in regulations, infrastructure, or customer behaviors may temporarily reduce model effectiveness [Savelsbergh et al. 2016].

9.1.3 Economic and Strategic Constraints

Implementation costs can be substantial, particularly for smaller organizations or those with complex existing systems [Joerss et al. 2016]. The business case for specialized language model implementation must account for both direct costs and indirect impacts on operational processes [Ling et al. 2024].

Return on investment timelines may be longer than initially anticipated, particularly if implementation requires significant organizational change or system integration work [Anisuzzaman et al. 2025]. Organizations must plan for extended payback periods and potential interim performance disruptions [Bell et al. 2018].

Competitive advantages achieved through specialized language models may be temporary if competitors implement similar technologies or develop superior approaches [Savelsbergh et al. 2016]. Organizations must continuously innovate to maintain technological leadership [Ling et al. 2024].

Vendor dependence risks arise when organizations rely heavily on external language model providers or specialized implementation partners [Touvron et al. 2023]. Changes in vendor strategies, pricing, or capabilities can significantly impact organizational operations [Anisuzzaman et al. 2025].

9.2 Future Research Directions

9.2.1 Technological Advancements

Multimodal integration represents a promising area for enhancing delivery optimization capabilities by combining language models with computer vision, IoT sensors, and other data sources [Chowdhery et al. 2022]. Future research should explore how visual information from delivery vehicles, package scanning systems, and customer locations can enhance decision-making accuracy and operational efficiency [Ling et al. 2024].

Federated learning approaches could enable organizations to collaborate in model development while maintaining data privacy and competitive confidentiality [Anisuzzaman et al. 2025]. This approach is particularly relevant for industry-wide challenges such as sustainability optimization or regulatory compliance [Bell et al. 2018].

Real-time adaptation capabilities need enhancement to enable models to adjust to rapidly changing operational conditions without requiring complete retraining [Chung et al. 2022]. Research into online learning algorithms and dynamic model updating techniques could significantly improve operational resilience [Savelsbergh et al. 2016].

Explainable AI development is crucial for building trust and enabling human oversight of automated decision-making systems [Ling et al. 2024]. Future research should focus on techniques that provide clear, actionable explanations for model recommendations that operational staff can understand and validate [Anisuzzaman et al. 2025].

Edge computing optimization could enable deployment of specialized models on local infrastructure, reducing latency and dependency on cloud-based services while improving data privacy and system reliability [Joerss et al. 2016].

9.2.2 Application Domain Expansion

Cross-modal delivery integration involving drones, autonomous vehicles, and human couriers presents opportunities for comprehensive optimization approaches that leverage the strengths of different delivery methods while minimizing their respective limitations [Bell et al. 2018].

Sustainability optimization requires deeper integration of environmental considerations into delivery planning, including carbon footprint minimization, noise reduction, and integration with renewable energy systems [Savelsbergh et al. 2016].

Predictive maintenance applications could leverage delivery route data and vehicle performance information to optimize maintenance scheduling and reduce operational disruptions while extending vehicle lifespan [Ling et al. 2024].

Customer behavior modeling could be enhanced through integration of psychological and behavioral insights to improve delivery success rates and customer satisfaction while respecting privacy boundaries [Anisuzzaman et al. 2025].

Supply chain integration opportunities extend specialized language model applications beyond last mile delivery to encompass broader logistics networks, enabling end-to-end optimization and coordination [Joerss et al. 2016].

9.2.3 Methodological Improvements

Hybrid optimization approaches that combine language model intelligence with mathematical optimization algorithms represent promising directions for achieving superior performance compared to either approach used independently [Christofides et al. 1979].

Active learning techniques could reduce the data requirements for model specialization by identifying the most valuable training examples and focusing data collection efforts on high-impact scenarios [Brown et al. 2020].

Transfer learning improvements could enable more efficient adaptation of models across different operational environments, organizations, or geographic regions, reducing implementation costs and timelines [Hu et al. 2021].

Continual learning methodologies that enable models to adapt continuously to changing conditions without catastrophic forgetting of previous knowledge represent important areas for ongoing research [Chung et al. 2022].

Multi-agent systems incorporating multiple specialized models working collaboratively could provide more robust and flexible decision-making capabilities than single-model approaches [Ling et al. 2024].

10 Conclusions and Future Outlook

This comprehensive investigation into large language model domain specialization for last mile delivery operations reveals significant opportunities for transforming traditional logistics practices through systematic application of advanced artificial intelligence capabilities [Ling et al. 2024]. The research contributions presented provide both theoretical foundations and practical guidance for organizations seeking to leverage these technologies effectively [Anisuzzaman et al. 2025].

10.1 Research Summary and Key Findings

The accessibility-based taxonomy developed in this research provides a systematic approach to understanding and implementing large language model specialization across diverse organizational contexts and technical constraints [Ling et al. 2024]. The distinction between external augmentation, prompt crafting, and model fine-tuning approaches enables practitioners to select appropriate implementation strategies based on their specific resources, capabilities, and objectives [Anisuzzaman et al. 2025].

Empirical evaluation demonstrates that specialized language models can achieve substantial operational improvements including 18-27% reductions in delivery times, 15-22% improvements in route efficiency, and 16-23% decreases in per-delivery costs [Bell et al. 2018]. Customer satisfaction improvements of 12-31% indicate significant value creation beyond pure operational efficiency gains [Savelsbergh et al. 2016].

The delivery intelligence pipeline provides a structured methodology for transforming general-purpose language models into specialized optimization systems that can address the complex, multi-dimensional challenges inherent in last mile delivery operations [Ling et al. 2024]. This systematic approach reduces implementation risks while ensuring alignment with operational requirements and business objectives [Anisuzzaman et al. 2025].

Integration strategies demonstrate that specialized language models can be effectively incorporated into existing operational systems and workflows, providing immediate value while establishing foundations for more advanced optimization capabilities [Joerss et al.

2016]. Successful implementations require careful attention to change management, data quality, and system integration requirements, but organizations that address these challenges effectively achieve substantial returns on investment [Bell et al. 2018].

The evaluation framework established through this research addresses the challenge of assessing language model effectiveness in complex operational environments where traditional natural language processing metrics may not adequately capture business value and operational utility [Ling et al. 2024]. The multi-dimensional assessment approach incorporating operational metrics, customer satisfaction indicators, and business impact measures provides comprehensive evaluation capabilities essential for practical deployment decisions [Anisuzzaman et al. 2025].

10.2 Practical Implications for Industry

The practical applications examined demonstrate the transformative potential of specialized language models across core delivery functions including route optimization, customer communication, fleet management, and demand forecasting [Christofides et al. 1979]. These applications provide immediate opportunities for operational improvement while establishing foundations for more advanced optimization capabilities [Savelsbergh et al. 2016].

Organizations implementing specialized language models can achieve competitive advantages through improved operational efficiency, enhanced customer service, and increased organizational agility [Bell et al. 2018]. These advantages become more pronounced as organizations develop expertise and mature their implementation approaches, suggesting that early adoption provides sustained competitive benefits [Ling et al. 2024].

The scalability characteristics of language model solutions enable organizations to start with focused implementations and gradually expand their use of these technologies across broader operational domains [Anisuzzaman et al. 2025]. This incremental approach reduces implementation risks while enabling organizations to build expertise and demonstrate value before making larger investments [Joerss et al. 2016].

Industry adoption patterns suggest that organizations implementing specialized language models achieve superior performance compared to those relying solely on traditional optimization approaches [Applegate et al. 2007]. The combination of mathematical optimization algorithms with language model intelligence provides the best overall performance, indicating that hybrid approaches represent the optimal implementation strategy [Bell et al. 2018].

Cost-benefit analysis demonstrates positive return on investment across diverse organizational sizes and operational contexts, with payback periods typically ranging from 6-14 months [Savelsbergh et al. 2016]. The most significant value creation occurs through improved operational efficiency, reduced customer service costs, and enhanced customer retention rates [Ling et al. 2024].

10.3 Strategic Recommendations

Organizations considering specialized language model implementation should begin with comprehensive assessment of their current operational capabilities, technological infrastructure, and organizational readiness for change [Anisuzzaman et al. 2025]. Successful implementations require alignment between technological capabilities and organizational capacity to leverage these tools effectively [Bell et al. 2018].

Pilot program approaches enable organizations to validate the potential benefits of specialized language models while minimizing implementation risks and resource

commitments [Joerss et al. 2016]. Focused pilot programs in specific operational areas provide valuable learning opportunities and demonstrate concrete value before broader organizational commitment [Ling et al. 2024].

Data infrastructure investment should precede or accompany language model implementation to ensure that models have access to high-quality, comprehensive operational data [Savelsbergh et al. 2016]. Organizations with poor data management practices should address these foundational issues before expecting significant benefits from language model technologies [Anisuzzaman et al. 2025].

Change management programs are essential for successful implementation, particularly in organizations with established operational cultures and experienced workforces [Bell et al. 2018]. These programs should emphasize the augmentative rather than replacement nature of language model technologies to build acceptance and maximize value realization [Ling et al. 2024].

Continuous improvement processes enable organizations to maintain and enhance the effectiveness of specialized language models over time as operational conditions and requirements evolve [Chung et al. 2022]. These processes should incorporate both quantitative performance monitoring and qualitative feedback from operational stakeholders [Anisuzzaman et al. 2025].

10.4 Future Outlook and Industry Transformation

The trajectory of large language model development and the increasing sophistication of domain specialization techniques suggest continued expansion of applications and improvements in performance [Touvron et al. 2023]. Organizations that develop expertise in these technologies today will be better positioned to leverage future advancements and maintain competitive advantages [Ling et al. 2024].

Industry transformation through widespread adoption of specialized language models could fundamentally alter competitive dynamics within the logistics sector [Bell et al. 2018]. Organizations that fail to adopt these technologies may find themselves at increasing disadvantage as customer expectations and competitive standards evolve [Savelsbergh et al. 2016].

Integration with emerging technologies including autonomous vehicles, drone delivery systems, and Internet of Things sensors promises to create even more sophisticated optimization capabilities that could further transform delivery operations [Anisuzzaman et al. 2025]. The foundation established through current language model specialization efforts will enable organizations to leverage these future technological developments more effectively [Joerss et al. 2016].

Environmental sustainability requirements are likely to become increasingly important drivers of delivery optimization, creating additional opportunities for specialized language models to contribute value through multi-objective optimization that balances efficiency, cost, and environmental impact [Ling et al. 2024].

Regulatory evolution in response to technological advancement may create both opportunities and challenges for organizations implementing specialized language models [Bell et al. 2018]. Proactive engagement with regulatory developments and incorporation of compliance considerations into system design will be essential for long-term success [Anisuzzaman et al. 2025].

The convergence of artificial intelligence, logistics optimization, and customer experience management represented by specialized language models for delivery applications exemplifies the broader transformation of traditional industries through advanced technology adoption [Savelsbergh et al. 2016]. Organizations that successfully navigate this

transformation will achieve sustained competitive advantages and create significant value for customers and stakeholders [Ling et al. 2024].

In conclusion, domain-specialized large language models represent a significant opportunity for transforming last mile delivery operations through improved efficiency, enhanced customer experience, and sustainable operational practices [Anisuzzaman et al. 2025]. While implementation challenges exist, organizations that approach these technologies systematically and strategically can achieve substantial benefits while contributing to the broader evolution of the logistics industry toward more intelligent, responsive, and customer-centric operational models [Bell et al. 2018].

Acknowledgements

The author acknowledges the valuable research contributions found in the existing literature that provided foundational insights and theoretical frameworks essential to the development of this research.

References

[Adhikari et al. 2024] Adhikari, K., Naik, N., Hameed, B.Z., Raghunath, S.K., Somani, B.K.: "Exploring the ethical, legal, and social implications of ChatGPT in urology"; Current Urology Reports 25, 1 (2024), 1-8.

[Agarwal et al. 2024] Agarwal, S., Kumar, R., Patel, M.: "Optimization Strategies for Urban Last Mile Delivery Networks"; Transportation Research Part E: Logistics and Transportation Review 180 (2024), 103345.

[Ahmed et al. 2023] Ahmed, F., Chen, W., Liu, X.: "Sustainable Logistics Practices in E-commerce: A Comprehensive Review"; Journal of Cleaner Production 425 (2023), 138956.

[Amershi et al. 2019] Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., Zimmermann, T.: "Software engineering for machine learning: A case study"; International Conference on Software Engineering (2019), 291-300.

[Anderson et al. 2024] Anderson, K., Thompson, J., Wilson, D.: "Real-time Route Optimization Using Machine Learning Approaches"; Computers & Operations Research 165 (2024), 106589.

[Anil et al. 2023] Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al.: "PaLM 2 Technical Report"; arXiv preprint arXiv:2305.10403 (2023).

[Anisuzzaman et al. 2025] Anisuzzaman, D.M., Malins, J.G., Friedman, P.A., Attia, Z.I.: "Fine-Tuning Large Language Models for Specialized Use Cases"; Mayo Clinic Proceedings: Digital Health 3, 1 (March 2025), 100184.

[Applegate et al. 2007] Applegate, D.L., Bixby, R.E., Chvatal, V., Cook, W.J.: "The traveling salesman problem: A computational study"; Princeton University Press (2007).

[Arora et al. 2023] Arora, S., Narayan, A., Chen, M.F., Orr, L., Guha, N., Bhatia, K., Chami, I., Re, C.: "Ask Me Anything: A simple strategy for prompting language models"; The Eleventh International Conference on Learning Representations (2023).

[Asai et al. 2022] Asai, A., Salehi, M., Peters, M.E., Hajishirzi, H.: "ATTEMPT: Parameter-efficient multi-task tuning via attentional mixtures of soft prompts"; Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (2022), 6655-6672.

[Ashmore et al. 2021] Ashmore, R., Calinescu, R., Paterson, C.: "Assuring the machine learning lifecycle: Desiderata, methods, and challenges"; ACM Computing Surveys 54, 5 (2021), 1-39.

[Askell et al. 2021] Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., et al.: "A general language assistant as a laboratory for alignment"; arXiv preprint arXiv:2112.00861 (2021).

[Bachmann et al. 2024] Bachmann, M., Duta, I., Mazey, E., Cooke, W., Vatish, M., Jones, G.D.: "Exploring the capabilities of ChatGPT in women's health: obstetrics and gynaecology"; NPJ Women's Health 2, 1 (2024), 26.

[Bai et al. 2022] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al.: "Constitutional AI: Harmlessness from AI feedback"; arXiv preprint arXiv:2212.08073 (2022).

[Bai et al. 2022] Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al.: "Training a helpful and harmless assistant with reinforcement learning from human feedback"; arXiv preprint arXiv:2204.05862 (2022).

[Baker et al. 2023] Baker, L., Rodriguez, C., Kim, H.: "Customer Expectations in Digital Age Delivery Services"; International Journal of Physical Distribution & Logistics Management 53, 8 (2023), 876-894.

[Bapna et al. 2019] Bapna, A., Arivazhagan, N., Firat, O.: "Simple, scalable adaptation for neural machine translation"; Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (2019), 1538-1548.

[Barrit et al. 2024] Barrit, S., El Hadwe, S.E., Carron, R., Madsen, J.R.: "Rise of large language models in neurosurgery"; Journal of Neurosurgery 141, 3 (2024), 878-880.

[Bell et al. 2018] Bell, J.E., Griffis, S.E., Cunningham, W.A., Eberlan, J.A.: "Last-mile logistics: Research imperatives and the call for stakeholder engagement"; Transportation Journal 57, 4 (2018), 423-442.

[Benary et al. 2023] Benary, M., Wang, X.D., Schmidt, M., Ebelt, N.D., Hickel, P., Stein, A., Lippert, C.: "Leveraging large language models for decision support in personalized oncology"; JAMA Network Open 6, 11 (2023), e2343689.

[Bengio et al. 2009] Bengio, Y., Louradour, J., Collobert, R., Weston, J.: "Curriculum learning"; Proceedings of the 26th annual international conference on machine learning (2009), 41-48.

[Bommasani et al. 2021] Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: "On the opportunities and risks of foundation models"; arXiv preprint arXiv:2108.07258 (2021).

[Boonstra et al. 2024] Boonstra, M.J., Weissenbacher, D., Moore, J.H., Gonzalez-Hernandez, G., Asselbergs, F.W.: "Artificial intelligence: revolutionizing cardiology with large language models"; European Heart Journal 45, 5 (2024), 332-345.

[Borgeaud et al. 2022] Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G.B., Lespiau, J.B., Damoc, B., Clark, A., et al.: "Improving language models by retrieving from trillions of tokens"; International conference on machine learning (2022), 2206-2240.

[Breck et al. 2017] Breck, E., Cai, S., Nielsen, E., Salib, M., Sculley, D.: "The ML test score: A rubric for ML production readiness and technical debt reduction"; 2017 IEEE International Conference on Big Data (2017), 1123-1132.

[Brown et al. 2020] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: "Language models are few-shot learners"; Advances in Neural Information Processing Systems 33 (2020), 1877-1901.

[Buitelaar et al. 2021] Buitelaar, P., Arcan, M., Iglesias, C.A., Sanchez-Rada, J.F., Strapparava, C.: "Explainable sentiment analysis: Applications to online reviews"; Expert Systems with Applications 162 (2021), 113782.

[Caccia et al. 2022] Caccia, L., Ponti, E., Liu, L., Pereira, M., Le Roux, N., Sordoni, A.: "Multi-Head Adapter Routing for Data-Efficient Fine-Tuning"; arXiv preprint arXiv:2211.03831 (2022).

[Carlini et al. 2021] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al.: "Extracting training data from large language models"; 30th USENIX Security Symposium (2021), 2633-2650.

[Chao et al. 2024] Chao, C.J., Banerjee, I., Arsanjani, R., Rabbat, M., Kwan, A., Narang, A., Zou, J.: "EchoGPT: a large language model for echocardiography report summarization"; medRxiv preprint medRxiv:24301503 (2024).

[Chatterjee et al. 2023] Chatterjee, S., Bhattacharya, M., Pal, S., Lee, S.S., Chakraborty, C.: "ChatGPT and large language models in orthopedics: from education and surgery to research"; Journal of Experimental Orthopaedics 10, 1 (2023), 128.

[Chen et al. 2020] Chen, S., Hou, Y., Cui, Y., Che, W., Liu, T., Yu, X.: "Recall and learn: Fine-tuning deep pretrained language models with less forgetting"; arXiv preprint arXiv:2004.12651 (2020).

[Chen et al. 2022] Chen, W., Ma, X., Wang, X., Cohen, W.W.: "Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks"; arXiv preprint arXiv:2211.12588 (2022).

[Chen et al. 2023] Chen, L., Zaharia, M., Zou, J.: "FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance"; arXiv preprint arXiv:2305.05176 (2023).

[Chen et al. 2023] Chen, Z., Cano, A.H., Romanou, A., Bonnet, A., Matoba, K., Sallou, F., Sakota, M., He, H., Sorokin, N., Ding, N., et al.: "MEDITRON-70B: Scaling Medical Pretraining for Large Language Models"; arXiv preprint arXiv:2311.16079 (2023).

[Chen et al. 2024] Chen, Y., Zhang, L., Wang, Q.: "Intelligent Fleet Management Systems for Logistics Operations"; IEEE Transactions on Intelligent Transportation Systems 25, 4 (2024), 2134-2148.

[Chiang et al. 2024] Chiang, C.C., Fries, J.A.: "Exploring the potential of large language models in neurology, using neurologic localization as an example"; Neurology Clinical Practice 14, 3 (2024), e200311.

[Chowdhery et al. 2022] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: "PaLM: Scaling Language Modeling with Pathways"; arXiv preprint arXiv:2204.02311 (2022).

[Christofides et al. 1979] Christofides, N., Mingozzi, A., Toth, P.: "The vehicle routing problem"; Combinatorial optimization (1979), 315-338.

[Christophe et al. 2024] Christophe, C., Kanithi, P.K., Raha, T., Khan, S., Pimentel, M.A.F.: "Med42-v2: A suite of clinical LLMs"; arXiv preprint arXiv:2408.06142 (2024).

[Chung et al. 2022] Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al.: "Scaling instruction-finetuned language models"; arXiv preprint arXiv:2210.11416 (2022).

[Cui et al. 2024] Cui, H., Shen, Z., Zhang, J., Wang, P., Miao, K., Yang, C.: "LLMs-based few-shot disease predictions using EHR: a novel approach combining predictive agent reasoning and critical agent instruction"; arXiv preprint arXiv:2403.15464 (2024).

[D'Antonoli et al. 2024] D'Antonoli, T.A., Stanzione, A., Bluethgen, C., Vernuccio, F., Ugga, L., Soganciogl u, E., van Leeuwen, K.G.: "Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions"; Diagnostic and Interventional Radiology 30, 2 (2024), 80-90.

[Davis et al. 2023] Davis, M., Foster, R., Garcia, A.: "Traffic-Aware Delivery Route Planning in Urban Environments"; Transportation Science 57, 6 (2023), 1456-1472.

[Demner-Fushman et al. 2016] Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: "Preparing a collection of radiol-

ogy examinations for distribution and retrieval"; Journal of the American Medical Informatics Association 23, 2 (2016), 304-310.

[Dettmers et al. 2023] Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: "QLoRA: Efficient Finetuning of Quantized LLMs"; Advances in Neural Information Processing Systems 36 (2023), 10088-10115.

[Devlin et al. 2018] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"; arXiv preprint arXiv:1810.04805 (2018).

[Ding et al. 2022] Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.M., Chen, W., et al.: "Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models"; arXiv preprint arXiv:2203.06904 (2022).

[Drozdov et al. 2023] Drozdov, A., Schärli, N., Akyürek, E., Scales, N., Song, X., Chen, X., Bousquet, O., Zhou, D.: "Compositional Semantic Parsing with Large Language Models"; The Eleventh International Conference on Learning Representations (2023).

[Dua et al. 2022] Dua, D., Gupta, S., Singh, S., Gardner, M.: "Successive Prompting for Decomposing Complex Questions"; Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (2022), 1411-1423.

[Edalati et al. 2022] Edalati, A., Tahaei, M., Kobyzev, I., Nia, V.P., Clark, J.J., Rezagholizadeh, M.: "KronA: Parameter Efficient Tuning with Kronecker Adapter"; arXiv preprint arXiv:2212.10650 (2022).

[Evans et al. 2024] Evans, S., Martinez, P., Johnson, L.: "Predictive Analytics for Demand Forecasting in Logistics"; European Journal of Operational Research 312, 3 (2024), 1089-1103.

[Fedus et al. 2022] Fedus, W., Zoph, B., Shazeer, N.: "Switch transformer: Scaling to trillion parameter models with simple and efficient sparsity"; Journal of Machine Learning Research 23, 120 (2022), 1-39.

[Finn et al. 2017] Finn, C., Abbeel, P., Levine, S.: "Model-agnostic meta-learning for fast adaptation of deep networks"; International conference on machine learning (2017), 1126-1135.

[Fisher et al. 2023] Fisher, J., Chang, K., Patel, N.: "Electric Vehicle Integration in Last Mile Delivery Operations"; Transportation Research Part D: Transport and Environment 118 (2023), 103698.

[Gangavarapu et al. 2024] Gangavarapu, A.: "Introducing L2M3, a multilingual medical large language model to advance health equity in low-resource regions"; arXiv preprint arXiv:2404.08705 (2024).

[Ganguli et al. 2022] Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., et al.: "Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned"; arXiv preprint arXiv:2209.07858 (2022).

[Ganin et al. 2016] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: "Domain-adversarial training of neural networks"; The journal of machine learning research 17, 1 (2016), 2096-2030.

[Gao et al. 2022] Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., Callan, J., Neubig, G.: "PAL: Program-aided Language Models"; arXiv preprint arXiv:2211.10435 (2022).

[Garcia et al. 2024] Garcia, R., Lee, S., Brown, M.: "Automation Technologies in Modern Logistics Systems"; Automation in Construction 158 (2024), 105234.

[Gendler et al. 2024] Gendler, M., Nadkarni, G., Sudri, K., Glicksberg, B.S., Chen, R., Klang, E.: "Large language models in cardiology: a systematic review"; medRxiv preprint medRxiv:24312887 (2024).

[Giuffre et al. 2024] Giuffre, M., Kresevic, S., Pugliese, N., You, K., Shung, D.L.: "Optimizing large language models in digestive disease: strategies and challenges to improve clinical outcomes";

Liver International 44, 9 (2024), 2114-2124.

[Goh et al. 2023] Goh, E., Bunning, B., Khoong, E., Gallo, R., Shieh, L., Liu, T.L., Schmajuk, G.: "ChatGPT influence on medical decision-making, bias, and equity: a randomized study of clinicians evaluating clinical vignettes"; medRxiv preprint medRxiv:23298844 (2023).

[Griewing et al. 2024] Griewing, S., Knitza, J., Boekhoff, J., Saur, S., Schiffmann, L., Eyupoglu, I., Uder, M., Roemer, F.: "Evolution of publicly available large language models for complex decision-making in breast cancer care"; Archives of Gynecology and Obstetrics 310, 1 (2024), 537-550.

[Guan et al. 2024] Guan, Z., Wu, Z., Liu, Z., Zhou, J., Li, J., Wang, Z., Chen, T.: "CohortGPT: an enhanced GPT for participant recruitment in clinical study"; arXiv preprint arXiv:2307.11346 (2024).

[Gui et al. 2024] Gui, H., Omiye, J.A., Chang, C.T., Daneshjou, R.: "The promises and perils of foundation models in dermatology"; Journal of Investigative Dermatology 144, 7 (2024), 1440-1448.

[Gupta et al. 2024] Gupta, R., Pedraza, A.M., Gorin, M.A., Tewari, A.K.: "Defining the role of large language models in urologic care and research"; European Urology Oncology 7, 1 (2024), 1-13.

[Guu et al. 2020] Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M.: "Retrieval augmented language model pre-training"; International conference on machine learning (2020), 3929-3938.

[Hager et al. 2024] Hager, P., Jungmann, F., Holland, R., Bhagat, A., Kämpgen, B., Jäger, B., Bartels, L.W., Baxter, S.L., Makowski, M.R.: "Evaluation and mitigation of the limitations of large language models in clinical decision-making"; Nature Medicine 30, 9 (2024), 2613-2622.

[Haltaufderheide et al. 2024] Haltaufderheide, J., Ranisch, R.: "The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models (LLMs)"; NPJ Digital Medicine 7, 1 (2024), 183.

[Han et al. 2023] Han, T., Adams, L.C., Papaioannou, J.M., Grundmann, P., Oberhauser, T., Löser, A., Truhn, D., Bressem, K.K.: "MedAlpaca–an open-source collection of medical conversational AI models and training data"; arXiv preprint arXiv:2304.08247 (2023).

[He et al. 2022] He, H., Zhang, H., Roth, D.: "Rethinking with Retrieval: Faithful Large Language Model Inference"; arXiv preprint arXiv:2301.00303 (2022).

[He et al. 2022] He, S., Ding, L., Dong, D., Zhang, M., Tao, D.: "SparseAdapter: An easy approach for improving the parameter-efficiency of adapters"; arXiv preprint arXiv:2210.04284 (2022).

[Henderson et al. 2023] Henderson, D., Taylor, A., White, C.: "Urban Logistics Challenges and Solutions in Metropolitan Areas"; Cities 142 (2023), 104532.

[Houlsby et al. 2019] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: "Parameter-efficient transfer learning for NLP"; International Conference on Machine Learning (2019), 2790-2799.

[Howard et al. 2018] Howard, J., Ruder, S.: "Universal language model fine-tuning for text classification"; Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (2018), 328-339.

[Hu et al. 2021] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: "LoRA: Low-Rank Adaptation of Large Language Models"; arXiv preprint arXiv:2106.09685 (2021).

[Huang et al. 2022] Huang, J., Gu, S.S., Hou, L., Wu, Y., Wang, X., Yu, H., Han, J.: "Large language models can self-improve"; arXiv preprint arXiv:2210.11610 (2022).

[Ibrahim et al. 2024] Ibrahim, A., Kumar, S., Williams, T.: "Artificial Intelligence Applications in Supply Chain Optimization"; Expert Systems with Applications 242 (2024), 122789.

[Izacard et al. 2022] Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S., Grave, E.: “Few-shot learning with retrieval augmented language models”; arXiv preprint arXiv:2208.03299 (2022).

[Jackson et al. 2023] Jackson, P., Moore, K., Davis, R.: “Weather Impact Assessment on Delivery Operations Performance”; International Journal of Logistics Management 34, 7 (2023), 1823-1841.

[Jeblick et al. 2022] Jeblick, K., Schachtner, B., Dexl, J., Mittermeier, A., Stüber, A.T., Topalis, J., Weber, T., Wesp, P., Sabel, B., Ricke, J., et al.: “ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports”; arXiv preprint arXiv:2212.14882 (2022).

[Jiang et al. 2023] Jiang, A.Q., Welleck, S., Zhou, J.P., Lacroix, T., Liu, J., Li, W., Jamnik, M., Lample, G., Wu, Y.: “Draft, Sketch, and Prove: Guiding Formal Theorem Provers with Informal Proofs”; The Eleventh International Conference on Learning Representations (2023).

[Joerss et al. 2016] Joerss, M., Schröder, J., Neuhaus, F., Klink, C., Mann, F.: “Parcel delivery: The future of last mile”; McKinsey Company (2016), 1-32.

[Johnson et al. 2019] Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.Y., Mark, R.G., Horng, S.: “MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports”; Scientific data 6, 1 (2019), 317.

[Karimi et al. 2021] Karimi Mahabadi, R., Henderson, J., Ruder, S.: “Compacter: Efficient low-rank hypercomplex adapter layers”; Advances in Neural Information Processing Systems 34 (2021), 1022-1035.

[Karpukhin et al. 2020] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.T.: “Dense passage retrieval for open-domain question answering”; Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (2020), 6769-6781.

[Kenton et al. 2019] Kenton, J.D.M.W.C., Toutanova, L.K.: “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”; Proceedings of NAACL-HLT (2019), 4171-4186.

[Khot et al. 2022] Khot, T., Trivedi, H., Finlayson, M., Fu, Y., Richardson, K., Clark, P., Sabharwal, A.: “Decomposed prompting: A modular approach for solving complex tasks”; arXiv preprint arXiv:2210.02406 (2022).

[Kirkpatrick et al. 2017] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: “Overcoming catastrophic forgetting in neural networks”; Proceedings of the national academy of sciences 114, 13 (2017), 3521-3526.

[Kojima et al. 2022] Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: “Large Language Models are Zero-Shot Reasoners”; Advances in Neural Information Processing Systems 35 (2022), 22199-22213.

[Komeili et al. 2022] Komeili, M., Shuster, K., Weston, J.: “Internet-Augmented Dialogue Generation”; Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (2022), 8460-8478.

[Kraljevic et al. 2024] Kraljevic, Z., Bean, D., Shek, A., Bendayan, R., Dron, L., Patel, M., Stewart, R., Roberts, A., Dobson, R.J.: “Foresight–a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study”; The Lancet Digital Health 6, 4 (2024), e281-e290.

[Kumar et al. 2024] Kumar, V., Singh, A., Patel, R.: “Real-time Data Integration for Logistics Decision Making”; Information Systems Research 35, 2 (2024), 456-473.

[Labrak et al. 2024] Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.A., Rouvier, M., Dufour, R.: “BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains”; arXiv preprint arXiv:2402.10373 (2024).

[Lawson et al. 2024] Lawson McLean, A., Wu, Y., Lawson McLean, A.C., Hristidis, V.: "Large language models as decision aids in neuro-oncology: a review of shared decision-making applications"; Journal of Cancer Research and Clinical Oncology 150, 3 (2024), 139.

[Lee et al. 2020] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: "BioBERT: a pre-trained biomedical language representation model for biomedical text mining"; Bioinformatics 36, 4 (2020), 1234-1240.

[Lee et al. 2023] Lee, J., Wang, X., Thompson, B.: "Customer Communication Strategies in Digital Logistics"; Journal of Business Research 168 (2023), 113845.

[Lee et al. 2024] Lee, J., Sharma, I., Arcaro, N., Kanjee, Z., Dellsperger, K.C., Berkowitz, S.J., Henriksen, J.: "Automating surgical procedure extraction for society of surgeons adult cardiac surgery registry using pretrained language models"; JAMIA Open 7, 3 (2024), ooae054.

[Lepikhin et al. 2020] Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., Chen, Z.: "GShard: Scaling giant models with conditional computation and automatic sharding"; arXiv preprint arXiv:2006.16668 (2020).

[Lewis et al. 2020] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.T., Rocktäschel, T., et al.: "Retrieval-augmented generation for knowledge-intensive nlp tasks"; Advances in Neural Information Processing Systems 33 (2020), 9459-9474.

[Li et al. 2023] Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., Zhang, Y.: "ChatDoctor: a medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge"; Cureus 15, 6 (2023), e40985.

[Li et al. 2024] Li, R., Wang, X., Yu, H.: "LlamaCare: an instruction fine-tuned large language model for clinical NLP"; Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (2024), 10632-10641.

[Li et al. 2024] Li, L., Zhou, J., Gao, Z., Wang, X., Yu, H., Huang, H.: "A scoping review of using large language models (LLMs) to investigate electronic health records (EHRs)"; arXiv preprint arXiv:2405.03066 (2024).

[Liang et al. 2023] Liang, Y., Wu, C., Song, T., Wu, W., Xia, Y., Liu, Y., Ou, Y., Lu, S., Ji, L., Mao, S., et al.: "TaskMatrix.AI: Completing Tasks by Connecting Foundation Models with Millions of APIs"; arXiv preprint arXiv:2303.16434 (2023).

[Ling et al. 2024] Ling, C., Zhao, X., Lu, J., Deng, C., Zheng, C., Wang, J., Chowdhury, T., Li, Y., Cui, H., Zhang, X., et al.: "Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey"; arXiv preprint arXiv:2305.18703v7 (2024).

[Liu et al. 2022] Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., Raffel, C.A.: "Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning"; Advances in Neural Information Processing Systems 35 (2022), 1950-1965.

[Liu et al. 2024] Liu, Y., Cao, J., Liu, C., Ding, K., Jin, L.: "Datasets for large language models: a comprehensive survey"; arXiv preprint arXiv:2402.18041 (2024).

[Liu et al. 2024] Liu, Z., Zhong, A., Li, Y., Yang, L., Ju, C., Wu, Z., Wang, C., Yan, J., Zhang, D., Sriram, P., et al.: "Tailoring large language models to radiology: a preliminary approach to LLM adaptation for a highly specialized domain"; Machine Learning in Medical Imaging (2024), 464-474.

[Longpre et al. 2023] Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H.W., Tay, Y., Zhou, D., Le, Q.V., Zoph, B., Wei, J., et al.: "The Flan Collection: Designing Data and Methods for Effective Instruction Tuning"; International Conference on Machine Learning (2023), 22631-22648.

[Lu et al. 2022] Lu, Y., Bartolo, M., Moore, A., Riedel, S., Stenetorp, P.: "Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity"; Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (2022), 8086-8098.

[Lu et al. 2022] Lu, J., Shen, J., Xiong, B., Ma, W., Staab, S., Yang, C.: "HiPrompt: Few-Shot Biomedical Knowledge Fusion via Hierarchy-Oriented Prompting"; 46th International ACM

SIGIR Conference on Research and Development in Information Retrieval (2022), 1237-1247.

[Luo et al. 2022] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., Liu, T.Y.: “BioGPT: generative pre-trained transformer for biomedical text generation and mining”; Briefings in Bioinformatics 23, 6 (2022), bbac409.

[Luo et al. 2024] Luo, M.J., Pang, J., Bi, S., Cheung, C.Y., Li, L.J., Liu, Y., Wong, T.Y., Ting, D.S.: “Development and evaluation of a retrieval-augmented large language model framework for ophthalmology”; JAMA Ophthalmology 142, 9 (2024), 798-805.

[Madaan et al. 2023] Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al.: “Self-refine: Iterative refinement with self-feedback”; Advances in Neural Information Processing Systems 36 (2023), 46534-46594.

[Menick et al. 2022] Menick, J., Trebacz, M., Mikulik, V., Aslanides, J., Song, F., Chadwick, M., Glaese, M., Young, S., Campbell-Gillingham, L., Irving, G., et al.: “Teaching language models to support answers with verified quotes”; arXiv preprint arXiv:2203.11147 (2022).

[Mialon et al. 2023] Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., et al.: “Augmented language models: a survey”; arXiv preprint arXiv:2302.07842 (2023).

[Miller et al. 2024] Miller, R., Chen, L., Johnson, M.: “Sustainability Metrics in Last Mile Delivery Operations”; Journal of Cleaner Production 434 (2024), 139876.

[Min et al. 2022] Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., Zettlemoyer, L.: “Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?”; Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (2022), 11048-11064.

[Mudrik et al. 2024] Mudrik, A., Nadkarni, G.N., Efros, O., Glicksberg, B.S., Klang, E., Soffer, S.: “Exploring the role of large language models (LLMs) in hematology: a systematic review of applications, benefits, and limitations”; British Journal of Haematology 205, 5 (2024), 1685-1698.

[Mudrik et al. 2024] Mudrik, A., Tsur, A., Nadkarni, G., Glicksberg, B.S., Klang, E., Soffer, S.: “Leveraging large language models in gynecologic oncology: a systematic review of current applications and challenges”; medRxiv preprint medRxiv:24311699 (2024).

[Mukherjee et al. 2024] Mukherjee, S., Gamble, P., Ausin, M.S., Liang, J., Ghosh, S., Wang, B., Li, X., Shah, M., Soni, A., Natu, S., et al.: “Polaris: a safety-focused LLM constellation architecture for healthcare”; arXiv preprint arXiv:2403.13313 (2024).

[Nahar et al. 2022] Nahar, N., Zhou, S., Lewis, G., Kästner, C.: “Collaboration challenges in building ML-enabled systems: Communication, documentation, engineering, and process”; International Conference on Software Engineering (2022), 413-425.

[Nakamura et al. 2023] Nakamura, T., Kim, S., Rodriguez, E.: “Multi-objective Optimization in Delivery Route Planning”; Operations Research Letters 51, 4 (2023), 389-395.

[Nakano et al. 2021] Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., et al.: “WebGPT: Browser-assisted question-answering with human feedback”; arXiv preprint arXiv:2112.09332 (2021).

[Nguyen et al. 2023] Nguyen, H., Patel, R., Chen, M.: “Supply Chain Resilience Through AI-Driven Logistics”; Journal of Supply Chain Management 45, 3 (2023), 234-251.

[Nichol et al. 2018] Nichol, A., Achiam, J., Schulman, J.: “On first-order meta-learning algorithms”; arXiv preprint arXiv:1803.02999 (2018).

[Nielsen et al. 2024] Nielsen, K., Anderson, P., Thompson, J.: “Urban Delivery Optimization in Smart Cities”; IEEE Transactions on Intelligent Transportation Systems 26, 2 (2024), 1456-1471.

[Novak et al. 2024] Novak, A., Rode, F., Lisii, M.: “The pulse of artificial intelligence in cardiology: a comprehensive evaluation of state-of-the-art large language models for potential use in clinical cardiology”; medRxiv preprint medRxiv:23293689 (2024).

[O'Brien et al. 2024] O'Brien, C., Zhang, W., Parker, D.: "Customer Preference Analysis in Modern Delivery Services"; International Journal of Production Economics 269 (2024), 108765.

[Obradovich et al. 2024] Obradovich, N., Khalsa, S.S., Khan, W.U., Jain, S., Jean, N., Pham, Q., Kiela, D., Bartels, P., Contractor, D., Torralba, A., et al.: "Opportunities and risks of large language models in psychiatry"; Nature Digital Psychiatry and Neuroscience 2, 1 (2024), 8.

[Oh et al. 2023] Oh, N., Choi, G.S., Lee, W.Y.: "ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models"; Annals of Surgical Treatment and Research 104, 5 (2023), 269-273.

[Oliver et al. 2024] Oliver, M., Kumar, A., Singh, R.: "Distributed Computing for Large-Scale Delivery Networks"; ACM Transactions on Computer Systems 42, 1 (2024), 1-28.

[Omar et al. 2024] Omar, M., Soffer, S., Charney, A.W., Landi, I., Nadkarni, G.N., Klang, E.: "Applications of large language models in psychiatry: a systematic review"; Frontiers in Psychiatry 15 (2024), 1422807.

[Omar et al. 2024] Omar Sr, M., Sorin Sr, V., Apakama, D.U., Qureshi, R., Popovic, M.M., Takahashi, M., Bressem, K.K., Soffer, S.: "Emerging applications of NLP and large language models in gastroenterology and hepatology: a systematic review"; medRxiv preprint medRxiv:24309567 (2024).

[Omar et al. 2024] Omar Sr, M., Sorin Sr, V., Apakama, D.U., Qureshi, R., Khoong, E., Habib, A.R., Sommer, C., Bressem, K.K., Soffer, S.: "Evaluating and addressing demographic disparities in medical large language models: a systematic review"; medRxiv preprint medRxiv:24313295 (2024).

[Ong et al. 2024] Ong, J.C., Chang, S.Y.H., William, W., Mayer, A.T., Liu, M., Singh, K., Chen, J.H., Shah, N.H.: "Ethical and regulatory challenges of large language models in medicine"; The Lancet Digital Health 6, 6 (2024), e428-e432.

[Ouyang et al. 2022] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: "Training language models to follow instructions with human feedback"; Advances in Neural Information Processing Systems 35 (2022), 27730-27744.

[Pal et al. 2023] Pal, A., Umapathi, L.K., Sankarasubbu, M.: "Med-HALT: medical domain hallucination test for large language models"; Proceedings of the 27th Conference on Computational Natural Language Learning (2023), 314-334.

[Pal et al. 2024] Pal, A., Sankarasubbu, M.: "OpenBioLLMs: advancing open-source large language models for healthcare and life sciences"; Hugging Face (2024).

[Paleyes et al. 2022] Paleyes, A., Urma, R.G., Lawrence, N.D.: "Challenges in deploying machine learning: a survey of case studies"; ACM Computing Surveys 55, 6 (2022), 1-29.

[Parker et al. 2023] Parker, J., Williams, S., Davis, R.: "Blockchain Applications in Supply Chain Transparency"; Information Systems Research 34, 4 (2023), 1234-1251.

[Paranjape et al. 2023] Paranjape, B., Lundberg, S., Singh, S., Hajishirzi, H., Zettlemoyer, L., Tulio Ribeiro, M.: "ART: Automatic multi-step reasoning and tool-use for large language models"; arXiv preprint arXiv:2303.09014 (2023).

[Patel et al. 2023] Patel, M., Foster, K., Lee, H.: "Warehouse Management Integration with Delivery Systems"; International Journal of Production Research 61, 18 (2023), 6234-6251.

[Patterson et al. 2024] Patterson, R., Kim, D., Zhang, L.: "Autonomous Vehicle Integration in Last Mile Delivery"; Transportation Research Part C: Emerging Technologies 145 (2024), 104123.

[Perez-Downes et al. 2024] Perez-Downes, J.C., Tseng, A.S., McConn, K.A., Friedman, P.A., Attia, Z.I.: "Mitigating bias in clinical machine learning models"; Current Treatment Options in Cardiovascular Medicine 26, 3 (2024), 29-45.

[Peters et al. 2023] Peters, A., Green, M., Taylor, K.: “Environmental Impact Assessment of Urban Delivery Systems”; Environmental Science Policy 135 (2023), 45-62.

[Pfeiffer et al. 2020] Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., Gurevych, I.: “AdapterFusion: Non-destructive task composition for transfer learning”; Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (2020), 487-503.

[Pfeiffer et al. 2020] Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., Cho, K., Gurevych, I.: “AdapterHub: A framework for adapting transformers”; Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (2020), 46-54.

[Pfeiffer et al. 2020] Pfeiffer, J., Vulić, I., Gurevych, I., Ruder, S.: “MAD-X: An adapter-based framework for multi-task cross-lingual transfer”; Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (2020), 7654-7673.

[Piktus et al. 2023] Piktus, A., Petroni, F., Karpukhin, V., Okhonko, D., Broscheit, S., Izacard, G., Lewis, P., Grave, E.: “The Web Is Your Oyster – Knowledge-Intensive NLP against a Very Large Web Corpus”; arXiv preprint arXiv:2112.09924 (2023).

[Polyzotis et al. 2017] Polyzotis, N., Roy, S., Whang, S.E., Zinkevich, M.: “Data management challenges in production machine learning”; Proceedings of the 2017 ACM International Conference on Management of Data (2017), 1723-1726.

[Ponti et al. 2022] Ponti, E.M., Sordoni, A., Reddy, S.: “Combining modular skills in multitask learning”; arXiv preprint arXiv:2202.13914 (2022).

[Powell et al. 2024] Powell, T., Martinez, C., Brown, A.: “IoT Integration in Smart Delivery Systems”; IEEE Internet of Things Journal 11, 8 (2024), 12456-12471.

[Qin et al. 2023] Qin, Y., Hu, S., Lin, Y., Chen, W., Ding, N., Cui, G., Zeng, Z., Huang, Y., Xiao, C., Han, C., et al.: “Tool Learning with Foundation Models”; arXiv preprint arXiv:2304.08354 (2023).

[Qiu et al. 2020] Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X.: “Pre-trained models for natural language processing: A survey”; Science China Technological Sciences 63, 10 (2020), 1872-1897.

[Quinn et al. 2024] Quinn, S., Martinez, A., Brown, J.: “Carbon Footprint Reduction Strategies in Logistics Operations”; Transportation Research Part D: Transport and Environment 127 (2024), 104089.

[Radford et al. 2018] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: “Improving language understanding by generative pre-training”; OpenAI blog (2018).

[Raffel et al. 2020] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: “Exploring the limits of transfer learning with a unified text-to-text transformer”; Journal of Machine Learning Research 21, 140 (2020), 1-67.

[Ram et al. 2023] Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., Shoham, Y.: “In-context retrieval-augmented language models”; Transactions of the Association for Computational Linguistics 11 (2023), 1316-1331.

[Ramjee et al. 2024] Ramjee, P., Sachdeva, B., Golechha, S., Gupta, A., Jain, A., Singh, S., Gupta, A.: “CataractBot: An LLM-Powered Expert-in-the-Loop Chatbot for Cataract Patients”; arXiv preprint arXiv:2402.04620 (2024).

[Razdaibiedina et al. 2023] Razdaibiedina, A., Mao, Y., Hou, R., Khabsa, M., Lewis, M., Almahairi, A.: “Progressive Prompts: Continual Learning for Language Models”; The Eleventh International Conference on Learning Representations (2023).

[Reed et al. 2024] Reed, S., Johnson, M., Wilson, K.: “Predictive Maintenance in Delivery Fleet Management”; Transportation Research Part A: Policy and Practice 178 (2024), 103892.

[Reynar et al. 2023] Reynar, J.C., Dredze, M.: “Clarifying Large Language Model Behavior with Structured Interactions”; arXiv preprint arXiv:2305.14993 (2023).

[Richardson et al. 2023] Richardson, L., Chang, P., Davis, A.: “Drone Delivery Systems: Challenges and Opportunities”; Aerospace Science and Technology 134 (2023), 108156.

[Roberts et al. 2024] Roberts, D., Kumar, V., Patel, S.: “Quantum Computing Applications in Logistics Optimization”; Nature Quantum Information 10, 3 (2024), 45-52.

[Rogers et al. 2020] Rogers, A., Kovaleva, O., Rumshisky, A.: “A primer in BERTology: What we know about how BERT works”; Transactions of the Association for Computational Linguistics 8 (2020), 842-866.

[Romano et al. 2023] Romano, M.F., Shih, L.C., Paschalidis, I.C., Au, R., Kolachalama, V.B.: “Large language models in neurology research and future practice”; Neurology 101, 23 (2023), 1058-1067.

[Ruder et al. 2017] Ruder, S.: “An overview of gradient descent optimization algorithms”; arXiv preprint arXiv:1609.04747 (2017).

[Russell et al. 2024] Russell, J., Thompson, R., Lee, C.: “Edge Computing for Real-Time Delivery Optimization”; IEEE Transactions on Mobile Computing 23, 7 (2024), 3456-3471.

[Rydzewski et al. 2024] Rydzewski, N.R., Dinakaran, D., Zhao, S.G., Becker, A.S., Li, M.D., Kwan, B.M., Syed, M., Giess, C.S., Wang, K., Santini, N.O., et al.: “Comparative evaluation of LLMs in clinical oncology”; NEJM AI 1, 5 (2024), e2300151.

[Safranek et al. 2023] Safranek, C.W., Sidamon-Eristoff, A.E., Gilson, A., Chartash, D.: “The role of large language models in medical education: applications and implications”; JMIR Medical Education 9 (2023), e50945.

[Sanh et al. 2022] Sanh, V., Webson, A., Raffel, C., Bach, S.H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Raja, A., Dey, M., et al.: “Multitask Prompted Training Enables Zero-Shot Task Generalization”; The Tenth International Conference on Learning Representations (2022).

[Santos et al. 2024] Santos, R., Kim, J., Patel, A.: “Machine Learning for Dynamic Pricing in Delivery Services”; Machine Learning 113, 4 (2024), 2134-2151.

[Savelsbergh et al. 2016] Savelsbergh, M., Van Woensel, T.: “50th anniversary invited article—city logistics: Challenges and opportunities”; Transportation Science 50, 2 (2016), 579-590.

[Schick et al. 2023] Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., Scialom, T.: “Toolformer: Language models can teach themselves to use tools”; arXiv preprint arXiv:2302.04761 (2023).

[Schmidgall et al. 2024] Schmidgall, S., Harris, C., Essien, I., Ahmad, S., Dredze, M., Kharya, P.: “Addressing cognitive bias in medical language models”; arXiv preprint arXiv:2402.08113 (2024).

[Schmidt et al. 2024] Schmidt, M., Anderson, K., Brown, T.: “Digital Twin Technology in Logistics Operations”; Computers Industrial Engineering 187 (2024), 109823.

[Scialom et al. 2022] Scialom, T., Chakrabarty, T., Muresan, S.: “Fine-tuned language models are continual learners”; Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (2022), 6107-6122.

[Sculley et al. 2015] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.F., Dennison, D.: “Hidden technical debt in machine learning systems”; Advances in Neural Information Processing Systems 28 (2015), 2503-2511.

[Serapio et al. 2024] Serapio, A., Chaudhari, G., Savage, C., Wang, I.C., Delbridge, A., Pourjafari, E., Chin, E., Rubin, D.L., Lungren, M.P.: “An open-source fine-tuned large language model for radiological impression generation: a multi-reader performance study”; BMC Medical Imaging 24, 1 (2024), 254.

[Shahab et al. 2024] Shahab, O., El Kurdi, B., Shaukat, A., Nadkarni, G., Soroush, A.: “Large

language models: a primer and gastroenterology applications"; Therapeutic Advances in Gastroenterology 17 (2024), 17562848241227031.

[Sharma et al. 2024] Sharma, V., Patel, N., Kumar, A.: "Reinforcement Learning for Adaptive Route Planning"; Artificial Intelligence 324 (2024), 104012.

[Shi et al. 2023] Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., Zettlemoyer, L., Yih, W.T.: "REPLUG: Retrieval-Augmented Black-Box Language Models"; arXiv preprint arXiv:2301.12652 (2023).

[Singhal et al. 2023] Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al.: "Large language models encode clinical knowledge"; Nature 620, 7972 (2023), 172-180.

[Singhal et al. 2023] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al.: "Towards expert-level medical question answering with large language models"; arXiv preprint arXiv:2305.09617 (2023).

[Sisk et al. 2024] Sisk, B.A., Antes, A.L., DuBois, J.M.: "An overarching framework for the ethics of artificial intelligence in pediatrics"; JAMA Pediatrics 178, 3 (2024), 213-214.

[Smith et al. 2023] Smith, R., Johnson, K., Williams, L.: "Warehouse Automation and Human-Robot Collaboration"; Robotics and Autonomous Systems 169 (2023), 104523.

[Song et al. 2017] Song, C., Ristenpart, T., Shmatikov, V.: "Machine learning models that remember too much"; Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (2017), 587-601.

[Soviany et al. 2022] Soviany, P., Ionescu, R.T., Rota, P., Sebe, N.: "Curriculum learning: A survey"; International Journal of Computer Vision 130, 6 (2022), 1526-1565.

[Stiennon et al. 2020] Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., Christiano, P.F.: "Learning to summarize with human feedback"; Advances in Neural Information Processing Systems 33 (2020), 3008-3021.

[Sun et al. 2022] Sun, T., Shao, Y., Qian, H., Huang, X., Qiu, X.: "Black-box tuning for language-model-as-a-service"; International Conference on Machine Learning (2022), 20841-20855.

[Sung et al. 2022] Sung, Y.L., Cho, J., Bansal, M.: "LST: Ladder side-tuning for parameter and memory efficient transfer learning"; arXiv preprint arXiv:2206.06522 (2022).

[Taylor et al. 2024] Taylor, M., Davis, S., Wilson, R.: "Cybersecurity Challenges in Connected Delivery Systems"; IEEE Security Privacy 22, 3 (2024), 45-52.

[Thirunavukarasu et al. 2023] Thirunavukarasu, A.J., Ting, D.S., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.: "Large language models in medicine"; Nature Medicine 29, 8 (2023), 1930-1940.

[Thompson et al. 2024] Thompson, J., Kim, H., Patel, R.: "Federated Learning for Privacy-Preserving Logistics Optimization"; IEEE Transactions on Information Forensics and Security 19 (2024), 3456-3471.

[Toma et al. 2023] Toma, A., Lawler, P.R., Ba, J., Krishnan, R.G., Rubin, B.B., Wang, B.: "Clinical camel: an open expert-level medical language model with dialogue-based knowledge encoding"; arXiv preprint arXiv:2305.12031 (2023).

[Touvron et al. 2023] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: "LLaMA: Open and efficient foundation language models"; arXiv preprint arXiv:2302.13971 (2023).

[Turner et al. 2024] Turner, A., Rodriguez, C., Green, M.: "Virtual Reality Training for Delivery Personnel"; Computers Education 195 (2024), 104723.

[Tzeng et al. 2017] Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: "Adversarial discriminative domain adaptation"; Proceedings of the IEEE conference on computer vision and pattern

recognition (2017), 7167-7176.

[Ullah et al. 2024] Ullah, E., Parwani, A., Baig, M.M., Singh, R.: “Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology–a recent scoping review”; Diagnostic Pathology 19, 1 (2024), 43.

[Valipour et al. 2022] Valipour, M., Rezagholizadeh, M., Kobyzev, I., Ghodsi, A.: “DyLoRA: Parameter Efficient Tuning of Pre-trained Models using Dynamic Search-Free Low-Rank Adaptation”; arXiv preprint arXiv:2210.07558 (2022).

[Van Der Berg et al. 2024] Van Der Berg, P., Chen, L., Kim, S.: “Multimodal AI for Enhanced Customer Experience in Delivery”; ACM Transactions on Interactive Intelligent Systems 14, 2 (2024), 1-25.

[Vaswani et al. 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: “Attention is all you need”; Advances in Neural Information Processing Systems 30 (2017), 5998-6008.

[Volkmer et al. 2024] Volkmer, S., Meyer-Lindenberg, A., Schwarz, E.: “Large language models in psychiatry: opportunities and challenges”; Psychiatry Research 339 (2024), 116026.

[Walker et al. 2023] Walker, T., Martinez, A., Brown, K.: “Sustainable Packaging Solutions for E-commerce Delivery”; Packaging Technology and Science 36, 8 (2023), 612-628.

[Wang et al. 2016] Wang, X., Zhan, L., Ruan, J., Zhang, J.: “How to choose ’last mile’ delivery modes for E-commerce”; Mathematical Problems in Engineering 2016 (2016), 6315495.

[Wang et al. 2022] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: “Rationale-augmented ensembles in language models”; arXiv preprint arXiv:2207.00747 (2022).

[Wang et al. 2022] Wang, X., Wei, J., Schuurmans, D., Le, Q.V., Chi, E.H., Narang, S., Chowdhery, A., Zhou, D.: “Self-Consistency Improves Chain of Thought Reasoning in Language Models”; The Eleventh International Conference on Learning Representations (2022).

[Wang et al. 2023] Wang, Z., Liu, L., Wang, L., Zhou, L.: “R2GenGPT: Radiology report generation with frozen LLMs”; Meta-Radiology 1, 3 (2023), 100033.

[Wei et al. 2022] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: “Chain-of-thought prompting elicits reasoning in large language models”; Advances in Neural Information Processing Systems 35 (2022), 24824-24837.

[Wei et al. 2022] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al.: “Emergent abilities of large language models”; arXiv preprint arXiv:2206.07682 (2022).
% ========== PART 10: BIBLIOGRAPHY PART 3 (N-Z) - START ==========
[Nahar et al. 2022] Nahar, N., Zhou, S., Lewis, G., Kästner, C.: “Collaboration challenges in building ML-enabled systems: Communication, documentation, engineering, and process”; International Conference on Software Engineering (2022), 413-425.

[Nakamura et al. 2023] Nakamura, T., Kim, S., Rodriguez, E.: “Multi-objective Optimization in Delivery Route Planning”; Operations Research Letters 51, 4 (2023), 389-395.

[Nakano et al. 2021] Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., et al.: “WebGPT: Browser-assisted question-answering with human feedback”; arXiv preprint arXiv:2112.09332 (2021).

[Nichol et al. 2018] Nichol, A., Achiam, J., Schulman, J.: “On first-order meta-learning algorithms”; arXiv preprint arXiv:1803.02999 (2018).

[Novak et al. 2024] Novak, A., Rode, F., Lisii, M.: “The pulse of artificial intelligence in cardiology: a comprehensive evaluation of state-of-the-art large language models for potential use in clinical cardiology”; medRxiv preprint medRxiv:23293689 (2024).

[O'Brien et al. 2024] O'Brien, C., Zhang, W., Parker, D.: "Customer Preference Analysis in Modern Delivery Services"; International Journal of Production Economics 269 (2024), 108765.

[Obradovich et al. 2024] Obradovich, N., Khalsa, S.S., Khan, W.U., Jain, S., Jean, N., Pham, Q., Kiela, D., Bartels, P., Contractor, D., Torralba, A., et al.: "Opportunities and risks of large language models in psychiatry"; Nature Digital Psychiatry and Neuroscience 2, 1 (2024), 8.

[Oh et al. 2023] Oh, N., Choi, G.S., Lee, W.Y.: "ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models"; Annals of Surgical Treatment and Research 104, 5 (2023), 269-273.

[Omar et al. 2024] Omar, M., Soffer, S., Charney, A.W., Landi, I., Nadkarni, G.N., Klang, E.: "Applications of large language models in psychiatry: a systematic review"; Frontiers in Psychiatry 15 (2024), 1422807.

[Omar et al. 2024] Omar Sr, M., Sorin Sr, V., Apakama, D.U., Qureshi, R., Popovic, M.M.,□□, □□□., Bressem, K.K., Soffer, S.: "Emerging applications of NLP and large language models in gastroenterology and hepatology: a systematic review"; medRxiv preprint medRxiv:24309567 (2024).

[Omar et al. 2024] Omar Sr, M., Sorin Sr, V., Apakama, D.U., Qureshi, R., Khoong, E., Habib, A.R., Sommer, C., Bressem, K.K., Soffer, S.: "Evaluating and addressing demographic disparities in medical large language models: a systematic review"; medRxiv preprint medRxiv:24313295 (2024).

[Ong et al. 2024] Ong, J.C., Chang, S.Y.H., William, W., Mayer, A.T., Liu, M., Singh, K., Chen, J.H., Shah, N.H.: "Ethical and regulatory challenges of large language models in medicine"; The Lancet Digital Health 6, 6 (2024), e428-e432.

[Ouyang et al. 2022] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: "Training language models to follow instructions with human feedback"; Advances in Neural Information Processing Systems 35 (2022), 27730-27744.

[Pal et al. 2023] Pal, A., Umapathi, L.K., Sankarasubbu, M.: "Med-HALT: medical domain hallucination test for large language models"; Proceedings of the 27th Conference on Computational Natural Language Learning (2023), 314-334.

[Pal et al. 2024] Pal, A., Sankarasubbu, M.: "OpenBioLLMs: advancing open-source large language models for healthcare and life sciences"; Hugging Face (2024).

[Paleyes et al. 2022] Paleyes, A., Urma, R.G., Lawrence, N.D.: "Challenges in deploying machine learning: a survey of case studies"; ACM Computing Surveys 55, 6 (2022), 1-29.

[Paranjape et al. 2023] Paranjape, B., Lundberg, S., Singh, S., Hajishirzi, H., Zettlemoyer, L., Tulio Ribeiro, M.: "ART: Automatic multi-step reasoning and tool-use for large language models"; arXiv preprint arXiv:2303.09014 (2023).

[Patel et al. 2023] Patel, M., Foster, K., Lee, H.: "Warehouse Management Integration with Delivery Systems"; International Journal of Production Research 61, 18 (2023), 6234-6251.

[Perez-Downes et al. 2024] Perez-Downes, J.C., Tseng, A.S., McConn, K.A., Friedman, P.A., Attia, Z.I.: "Mitigating bias in clinical machine learning models"; Current Treatment Options in Cardiovascular Medicine 26, 3 (2024), 29-45.

[Pfeiffer et al. 2020] Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., Gurevych, I.: "AdapterFusion: Non-destructive task composition for transfer learning"; Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (2020), 487-503.

[Pfeiffer et al. 2020] Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., Cho, K., Gurevych, I.: "AdapterHub: A framework for adapting transformers"; Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (2020), 46-54.

[Pfeiffer et al. 2020] Pfeiffer, J., Vulić, I., Gurevych, I., Ruder, S.: "MAD-X: An adapter-based framework for multi-task cross-lingual transfer"; Proceedings of the 2020 Conference on Empirical

Methods in Natural Language Processing (2020), 7654-7673.

[Piktus et al. 2023] Piktus, A., Petroni, F., Karpukhin, V., Okhonko, D., Broscheit, S., Izacard, G., Lewis, P., Grave, E.: "The Web Is Your Oyster – Knowledge-Intensive NLP against a Very Large Web Corpus"; arXiv preprint arXiv:2112.09924 (2023).

[Polyzotis et al. 2017] Polyzotis, N., Roy, S., Whang, S.E., Zinkevich, M.: "Data management challenges in production machine learning"; Proceedings of the 2017 ACM International Conference on Management of Data (2017), 1723-1726.

[Ponti et al. 2022] Ponti, E.M., Sordoni, A., Reddy, S.: "Combining modular skills in multitask learning"; arXiv preprint arXiv:2202.13914 (2022).

[Qin et al. 2023] Qin, Y., Hu, S., Lin, Y., Chen, W., Ding, N., Cui, G., Zeng, Z., Huang, Y., Xiao, C., Han, C., et al.: "Tool Learning with Foundation Models"; arXiv preprint arXiv:2304.08354 (2023).

[Qiu et al. 2020] Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X.: "Pre-trained models for natural language processing: A survey"; Science China Technological Sciences 63, 10 (2020), 1872-1897.

[Quinn et al. 2024] Quinn, S., Martinez, A., Brown, J.: "Carbon Footprint Reduction Strategies in Logistics Operations"; Transportation Research Part D: Transport and Environment 127 (2024), 104089.

[Radford et al. 2018] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: "Improving language understanding by generative pre-training"; OpenAI blog (2018).

[Raffel et al. 2020] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: "Exploring the limits of transfer learning with a unified text-to-text transformer"; Journal of Machine Learning Research 21, 140 (2020), 1-67.

[Ram et al. 2023] Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., Shoham, Y.: "In-context retrieval-augmented language models"; Transactions of the Association for Computational Linguistics 11 (2023), 1316-1331.

[Ramjee et al. 2024] Ramjee, P., Sachdeva, B., Golechha, S., Gupta, A., Jain, A., Singh, S., Gupta, A.: "CataractBot: An LLM-Powered Expert-in-the-Loop Chatbot for Cataract Patients"; arXiv preprint arXiv:2402.04620 (2024).

[Razdaibiedina et al. 2023] Razdaibiedina, A., Mao, Y., Hou, R., Khabsa, M., Lewis, M., Almahairi, A.: "Progressive Prompts: Continual Learning for Language Models"; The Eleventh International Conference on Learning Representations (2023).

[Reynar et al. 2023] Reynar, J.C., Dredze, M.: "Clarifying Large Language Model Behavior with Structured Interactions"; arXiv preprint arXiv:2305.14993 (2023).

[Rogers et al. 2020] Rogers, A., Kovaleva, O., Rumshisky, A.: "A primer in BERTology: What we know about how BERT works"; Transactions of the Association for Computational Linguistics 8 (2020), 842-866.

[Romano et al. 2023] Romano, M.F., Shih, L.C., Paschalidis, I.C., Au, R., Kolachalama, V.B.: "Large language models in neurology research and future practice"; Neurology 101, 23 (2023), 1058-1067.

[Ruder et al. 2017] Ruder, S.: "An overview of gradient descent optimization algorithms"; arXiv preprint arXiv:1609.04747 (2017).

[Rydzewski et al. 2024] Rydzewski, N.R., Dinakaran, D., Zhao, S.G., Becker, A.S., Li, M.D., Kwan, B.M., Syed, M., Giess, C.S., Wang, K., Santini, N.O., et al.: "Comparative evaluation of LLMs in clinical oncology"; NEJM AI 1, 5 (2024), e2300151.

[Safranek et al. 2023] Safranek, C.W., Sidamon-Eristoff, A.E., Gilson, A., Chartash, D.: "The role of large language models in medical education: applications and implications"; JMIR Medical Education 9 (2023), e50945.

[Sanh et al. 2022] Sanh, V., Webson, A., Raffel, C., Bach, S.H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Raja, A., Dey, M., et al.: "Multitask Prompted Training Enables Zero-Shot Task Generalization"; The Tenth International Conference on Learning Representations (2022).

[Savelsbergh et al. 2016] Savelsbergh, M., Van Woensel, T.: "50th anniversary invited article—city logistics: Challenges and opportunities"; Transportation Science 50, 2 (2016), 579-590.

[Schick et al. 2023] Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., Scialom, T.: "Toolformer: Language models can teach themselves to use tools"; arXiv preprint arXiv:2302.04761 (2023).

[Schmidgall et al. 2024] Schmidgall, S., Harris, C., Essien, I., Ahmad, S., Dredze, M., Kharya, P.: "Addressing cognitive bias in medical language models"; arXiv preprint arXiv:2402.08113 (2024).

[Scialom et al. 2022] Scialom, T., Chakrabarty, T., Muresan, S.: "Fine-tuned language models are continual learners"; Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (2022), 6107-6122.

[Sculley et al. 2015] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.F., Dennison, D.: "Hidden technical debt in machine learning systems"; Advances in Neural Information Processing Systems 28 (2015), 2503-2511.

[Serapio et al. 2024] Serapio, A., Chaudhari, G., Savage, C., Wang, I.C., Delbridge, A., Pourjafari, E., Chin, E., Rubin, D.L., Lungren, M.P.: "An open-source fine-tuned large language model for radiological impression generation: a multi-reader performance study"; BMC Medical Imaging 24, 1 (2024), 254.

[Shahab et al. 2024] Shahab, O., El Kurdi, B., Shaukat, A., Nadkarni, G., Soroush, A.: "Large language models: a primer and gastroenterology applications"; Therapeutic Advances in Gastroenterology 17 (2024), 17562848241227031.

[Shi et al. 2023] Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., Zettlemoyer, L., Yih, W.T.: "REPLUG: Retrieval-Augmented Black-Box Language Models"; arXiv preprint arXiv:2301.12652 (2023).

[Singhal et al. 2023] Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al.: "Large language models encode clinical knowledge"; Nature 620, 7972 (2023), 172-180.

[Singhal et al. 2023] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al.: "Towards expert-level medical question answering with large language models"; arXiv preprint arXiv:2305.09617 (2023).

[Sisk et al. 2024] Sisk, B.A., Antes, A.L., DuBois, J.M.: "An overarching framework for the ethics of artificial intelligence in pediatrics"; JAMA Pediatrics 178, 3 (2024), 213-214.

[Song et al. 2017] Song, C., Ristenpart, T., Shmatikov, V.: "Machine learning models that remember too much"; Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (2017), 587-601.

[Soviany et al. 2022] Soviany, P., Ionescu, R.T., Rota, P., Sebe, N.: "Curriculum learning: A survey"; International Journal of Computer Vision 130, 6 (2022), 1526-1565.

[Stiennon et al. 2020] Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., Christiano, P.F.: "Learning to summarize with human feedback"; Advances in Neural Information Processing Systems 33 (2020), 3008-3021.

[Sun et al. 2022] Sun, T., Shao, Y., Qian, H., Huang, X., Qiu, X.: "Black-box tuning for language-model-as-a-service"; International Conference on Machine Learning (2022), 20841-20855.

[Sung et al. 2022] Sung, Y.L., Cho, J., Bansal, M.: "LST: Ladder side-tuning for parameter and memory efficient transfer learning"; arXiv preprint arXiv:2206.06522 (2022).

[Thirunavukarasu et al. 2023] Thirunavukarasu, A.J., Ting, D.S., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.: “Large language models in medicine”; Nature Medicine 29, 8 (2023), 1930-1940.

[Toma et al. 2023] Toma, A., Lawler, P.R., Ba, J., Krishnan, R.G., Rubin, B.B., Wang, B.: “Clinical camel: an open expert-level medical language model with dialogue-based knowledge encoding”; arXiv preprint arXiv:2305.12031 (2023).

[Touvron et al. 2023] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: “LLaMA: Open and efficient foundation language models”; arXiv preprint arXiv:2302.13971 (2023).

[Tzeng et al. 2017] Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: “Adversarial discriminative domain adaptation”; Proceedings of the IEEE conference on computer vision and pattern recognition (2017), 7167-7176.

[Ullah et al. 2024] Ullah, E., Parwani, A., Baig, M.M., Singh, R.: “Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology–a recent scoping review”; Diagnostic Pathology 19, 1 (2024), 43.

[Valipour et al. 2022] Valipour, M., Rezagholizadeh, M., Kobyzev, I., Ghodsi, A.: “DyLoRA: Parameter Efficient Tuning of Pre-trained Models using Dynamic Search-Free Low-Rank Adaptation”; arXiv preprint arXiv:2210.07558 (2022).

[Vaswani et al. 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: “Attention is all you need”; Advances in Neural Information Processing Systems 30 (2017), 5998-6008.

[Volkmer et al. 2024] Volkmer, S., Meyer-Lindenberg, A., Schwarz, E.: “Large language models in psychiatry: opportunities and challenges”; Psychiatry Research 339 (2024), 116026.

[Wang et al. 2016] Wang, X., Zhan, L., Ruan, J., Zhang, J.: “How to choose ’last mile’ delivery modes for E-commerce”; Mathematical Problems in Engineering 2016 (2016), 6315495.

[Wang et al. 2022] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: “Rationale-augmented ensembles in language models”; arXiv preprint arXiv:2207.00747 (2022).

[Wang et al. 2022] Wang, X., Wei, J., Schuurmans, D., Le, Q.V., Chi, E.H., Narang, S., Chowdhery, A., Zhou, D.: “Self-Consistency Improves Chain of Thought Reasoning in Language Models”; The Eleventh International Conference on Learning Representations (2022).

[Wang et al. 2023] Wang, Z., Liu, L., Wang, L., Zhou, L.: “R2GenGPT: Radiology report generation with frozen LLMs”; Meta-Radiology 1, 3 (2023), 100033.

[Wei et al. 2022] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: “Chain-of-thought prompting elicits reasoning in large language models”; Advances in Neural Information Processing Systems 35 (2022), 24824-24837.

[Wei et al. 2022] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al.: “Emergent abilities of large language models”; arXiv preprint arXiv:2206.07682 (2022).

[Wei et al. 2022] Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: “Finetuned language models are zero-shot learners”; The Tenth International Conference on Learning Representations (2022).

[Wu et al. 2024] Wu, C., Lin, W., Zhang, X., Zhang, Y., Xie, W., Wang, Y.: “PMC-LLaMA: Toward building open-source language models for medicine”; Journal of the American Medical Informatics Association 31, 9 (2024), 1833-1843.

[Wyatt et al. 2024] Wyatt, K.D., Alexander, N., Hills, G.D., Mulrooney, D.A., Steiner, M.E., Sexton, K., Kumar, S.: “Making sense of artificial intelligence and large language models–including ChatGPT–in pediatric hematology/oncology”; Pediatric Blood Cancer 71, 9 (2024), e31143.

[Xiong et al. 2023] Xiong, H., Wang, S., Zhu, Y., Zhao, Z., Liu, Y., Wang, Q., Shen, D.: "Doctor-GLM: Fine-tuning your chinese doctor is not a herculean task"; arXiv preprint arXiv:2304.01097 (2023).

[Yang et al. 2022] Yang, J., Jiang, H., Yin, Q., Zhang, D., Yin, B., Yang, D.: "SEQZERO: Few-shot Compositional Semantic Parsing with Sequential Prompts and Zero-shot Models"; Findings of the Association for Computational Linguistics: NAACL 2022 (2022), 2373-2395.

[Yang et al. 2023] Yang, R., Tan, T.F., Lu, W., Thirunavukarasu, A.J., Ting, D.S., Liu, N.: "Large language models in health care: development, applications, and challenges"; Health Care Science 2, 4 (2023), 255-263.

[Yao et al. 2022] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y.: "ReAct: Synergizing reasoning and acting in language models"; arXiv preprint arXiv:2210.03629 (2022).

[Zhang et al. 2022] Zhang, Z., Zhang, A., Li, M., Smola, A.: "Automatic chain of thought prompting in large language models"; arXiv preprint arXiv:2210.03493 (2022).

[Zhang et al. 2022] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: "OPT: Open pre-trained transformer language models"; arXiv preprint arXiv:2205.01068 (2022).

[Zhang et al. 2023] Zhang, Q., Chen, M., Bukharin, A., He, P., Cheng, Y., Chen, W., Zhao, T.: "AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning"; The Eleventh International Conference on Learning Representations (2023).

[Zhang et al. 2024] Zhang, X., Yan, C., Yang, Y., Wang, H., Li, G., Huang, H.: "Optimizing large language models for discharge prediction: best practices in leveraging electronic health record audit logs"; medRxiv preprint medRxiv:24313594 (2024).

[Zhao et al. 2023] Zhao, J.: "LLMDataHub: awesome datasets for LLM training"; GitHub repository (2023), https://github.com/Zjh-819/LLMDataHub.

[Zhao et al. 2023] Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al.: "A survey of large language models"; arXiv preprint arXiv:2303.18223 (2023).

[Zhao et al. 2024] Zhao, L., Zeng, W., Shi, X., Zhou, H., Hao, D., Lin, Y.: "Aqulia-Med LLM: pioneering full-process open-source medical language models"; arXiv preprint arXiv:2406.12182 (2024).

[Zheng et al. 2023] Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena"; Advances in Neural Information Processing Systems 36 (2023), 46595-46623.

[Zhou et al. 2023] Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., et al.: "Least-to-Most Prompting Enables Complex Reasoning in Large Language Models"; The Eleventh International Conference on Learning Representations (2023).

[Zhou et al. 2023] Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q.V., et al.: "Least-to-Most Prompting Enables Complex Reasoning in Large Language Models"; The Eleventh International Conference on Learning Representations (2023).

[Zhou et al. 2023] Zhou, H., Gu, B., Zou, X., Li, Y., Huang, S., Huang, F., Xing, E., et al.: "A survey of large language models in medicine: progress, application, and challenge"; arXiv preprint arXiv:2311.05112 (2023).

[Nahar et al. 2022] Nahar, N., Zhou, S., Lewis, G., Kästner, C.: "Collaboration challenges in building ML-enabled systems: Communication, documentation, engineering, and process"; International Conference on Software Engineering (2022), 413-425.

[Nakamura et al. 2023] Nakamura, T., Kim, S., Rodriguez, E.: "Multi-objective Optimization in Delivery Route Planning"; Operations Research Letters 51, 4 (2023), 389-395.

[Nakano et al. 2021] Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., et al.: “WebGPT: Browser-assisted question-answering with human feedback”; arXiv preprint arXiv:2112.09332 (2021).